Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

What do officers learn from an AI-simulated suspect? An exploratory pre/post pilot of police questioning for a suspect with developmental language disorder

Christopher Baker

c.baker@qub.ac.uk

Queen's University Belfast

Fiona Stuart

Queen's University Belfast

Karen Rafferty

Queen's University Belfast

John Taggart

Queen's University Belfast

Research Article

Keywords: large language models, simulation-based training, investigative interviewing, developmental language disorder, speech, language and communication needs, automated assessment, human–AI interaction, police training

Posted Date: October 30th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11552083/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

What do officers learn from an AI-simulated suspect? An exploratory pre/post pilot of police questioning for a suspect with developmental language disorder

Christopher Baker (corresponding author)
c.baker@qub.ac.uk
School of Electronics, Electrical Engineering and Computer Science
Queen's University Belfast, Belfast, United Kingdom

Fiona Stuart
fstuart02@qub.ac.uk
School of Law
Queen's University Belfast, Belfast, United Kingdom

Karen Rafferty
k.rafferty@qub.ac.uk
School of Electronics, Electrical Engineering and Computer Science
Queen's University Belfast, Belfast, United Kingdom

John Taggart
john.taggart@qub.ac.uk
School of Law
Queen's University Belfast, Belfast, United Kingdom

Abstract

Background

People with speech, language and communication needs, including undiagnosed developmental language disorder (DLD), are over-represented in the criminal justice system, yet police questioning is rarely adapted to them. Large language models make simulated interview practice possible, but the effects of such training, and the accuracy of the automated scoring that supports it, are largely unknown. We evaluated the Forensic Interview Trainer, in which a simulated 19-year-old suspect with undiagnosed DLD responds through a large language model constrained by deterministic rules, and a rule-based auditor scores each officer's question against five communication strategies.

Results

Eighteen serving police officers interviewed the simulated suspect for five minutes, viewed a short briefing, and interviewed him again, and 927 questions were compared after three systematic errors in the auditor had been corrected. The proportion of questions classified as Recommended rose from 28.2% to 46.1% (mean within-officer change 18.0 percentage points, 95% confidence interval 8.1 to 31.0), and 15 of 18 officers improved, with fewer closed questions, difficult words and multi-part questions, and fewer content words per question. The original auditor misclassified 22% of questions, always under-rating the officer, and because its flags also drove the simulated suspect, 12% of questions received an automatic refusal triggered by an everyday word, a name or a transcription error. In exploratory analyses, questions opening with fillers such as "okay" fell from 51.1% to 17.6%, although the briefing never mentioned them. Officers rated the training highly, and several noticed the scoring errors, yet all agreed that the feedback was useful. When the suspect gave a non-answer, officers tended to narrow to closed questions or to challenge him rather than simplify.

Conclusions

A short briefing combined with practice against a language model-simulated suspect was associated with a large improvement in officers' questioning of a suspect with communication needs. However, errors in the automated scoring shaped both the simulated suspect's replies and the scores officers were shown, and the uncontrolled design cannot isolate the contribution of the simulation. Automated scoring in AI-mediated training should be validated against expert judgement before its feedback is used to teach.

Keywords

large language models; simulation-based training; investigative interviewing; developmental language disorder; speech, language and communication needs; automated assessment; human–AI interaction; police training

Background

Developmental language disorder (DLD), the term recommended for language disorder with no known differentiating condition (1), affects an estimated 7.6% of children at school entry (2), and it is markedly more common among those who come into contact with the criminal justice system. Youth offenders consistently perform below age-matched peers on measures of receptive and expressive language (3), none of a randomly selected sample of juvenile offenders reached their age-equivalent level of receptive vocabulary (4), and the language impairments of juvenile and adult offenders frequently pass through police and court procedures without anyone recognising them (5). Young adults whose DLD was identified and supported at school, by contrast, reported less contact with the police than their peers (6), which suggests that identification and support may matter as much as the disorder itself in this context. The investigative interview places heavy demands on precisely the abilities that DLD affects, since even the police caution is understood in full by only a minority of the general population, and less well still by those of below-average intellectual ability (7), while a suspect who misunderstands the questions put to them may give unreliable answers and may be at increased risk of making a false confession (5).

Best-practice guidance on questioning vulnerable witnesses and defendants is consistent in what it asks of the questioner: to use common words and plain language in place of legal terminology, to ask short questions containing one idea at a time, and to avoid negatives and tag questions, which are linguistically complex and, in the case of tags, powerfully suggestive (8). Experimental work supports these recommendations, since questions containing negatives, double negatives, several parts or difficult vocabulary confuse children, adolescents and young adults alike (9) and reduce the accuracy of adult witnesses, who rarely ask for a confusing question to be explained (10). In Northern Ireland, Registered Intermediaries can assist suspects with significant communication deficits during police interview (11), but their involvement depends on those difficulties first being recognised, and where they are not, it falls to the interviewing officer to adapt their own questioning. Field evidence suggests that such adaptation cannot be assumed: in recorded police interviews with suspects, fewer than 1% of questions were open-ended, although open questions drew the longest responses, while around two-fifths were closed yes/no questions (12).

Recommended questioning practices are not reliably acquired from guidance or classroom instruction alone, and field studies of child witness interviewing have shown that gains made under structured supervision are lost once feedback stops: in the work of Lamb and colleagues (13), investigators asked fewer open-ended and more option-posing and suggestive questions after supervision ended, and children's recalled detail fell accordingly. This has encouraged interest in simulation, in which trainees can rehearse questioning repeatedly against a consistent interviewee and receive feedback without risk to a real witness or suspect, and Pompedda, Zappalà and Santtila (14) found that trainees interviewing child avatars who received feedback on question types and on the accuracy of their conclusions used more open-ended and fewer closed questions than those who did not, within roughly an hour of practice. Large language models have since been used to drive such simulations in place of pre-scripted responses: Røed et al. (15) built a child avatar chatbot on a fine-tuned GPT-3 model, coupled it with an automatic classifier of question types that supplied direct feedback, and reported that the feedback group increased its use of open-ended questions while neither group reduced its use of closed questions, and comparable systems have been developed as simulated patients for history-

taking practice in medical education (16). These systems raise two difficulties that bear directly on training for communication with vulnerable people. The first is that a language model asked to play a character is performing a role rather than possessing the characteristics described (17), so it may depart from the profile it has been given; Holderried et al. (16) found that although nearly all of their simulated patient's answers were judged medically plausible, over half of the answers to questions the case script did not cover contained invented details, and some answers abandoned the patient role altogether. The second is that the automated scoring on which feedback depends is seldom validated, and a recent scoping review of language model-based virtual patients found that only three of 23 evaluations used validated assessment tools and only five measured learning outcomes objectively (18). We are not aware of any published evaluation of a language model-simulated suspect with a communication vulnerability, nor of one that checks whether the automated feedback given to interviewers faithfully reflects the best-practice guidance it is meant to apply.

The Forensic Interview Trainer (FIT) was developed to address these gaps for suspect interviewing. It presents officers with Jordan, a simulated 19-year-old suspect with undiagnosed DLD. Jordan is voiced and animated, and his replies are generated by a large language model from a persona profile and then constrained by a rule-based enforcement layer, so that he responds to poorly constructed questions in the way the profile describes rather than as the model might otherwise choose. Each officer question is also scored by a rule-based linguistic auditor, which flags five features that best-practice guidance on questioning people with communication needs advises against: closed questions, questions with a high number of information-carrying words (ICWs), negatively phrased questions, tag questions and difficult vocabulary (8). The auditor's flags are used in two ways. During the interview they drive the enforcement layer, and after the session they form the feedback given to the officer. Officers could therefore experience the consequences of their questioning in Jordan's replies and then see it summarised question by question.

This exploratory pilot study evaluated FIT with police officers, who each conducted a brief baseline interview, viewed a short presentation on the recommended strategies, conducted a second interview, and then saw the auditor's scores for both interviews. The study had two research questions: whether the proportion of questions consistent with recommended practice increased after training, and which of the recommended strategies showed the greatest change. Because the automated scoring was itself an untested component, shaping both the simulated suspect's replies and the feedback officers received, we also examined whether the auditor's classifications matched the best-practice guidance it was designed to implement. Where they did not, we corrected the auditor's rules and reported outcomes under both the original and corrected scoring. Finally, we exploredhow Jordan responded to different question types, how officers adapted their questioning, and how they viewed the training.

Methods

Design

The study was a single-arm pre/post pilot with no control group. Each officer interviewed the same simulated suspect twice, for five minutes each time, timed by the researcher, with a short briefing on questioning people with communication needs between the two interviews. Officers were shown the automated scoring of their questions only after the second interview. Because there was no control condition, any change between interviews cannot be attributed to the simulated suspect rather than to practice, familiarity with the scenario or the briefing, and the analysis is therefore exploratory, estimating the direction and size of change rather than testing the effectiveness of the training. The analysis plan was recorded before outcomes were compared between interviews, and analyses added after that point are identified as exploratory throughout. One correction to the automated scoring, for open prompts beginning "tell", "explain" or "describe", was added after the main analyses had first been run, following review by a co-author. On its own, it changed the classification of seven of the 927 questions.

Participants

Twenty-two sessions were run between 4 August and 4 September 2026. The first three were run with non-police volunteers, and the interview sessions of one officer were not recorded, so 18 serving police officers from the Police Service of Northern Ireland, interviewed between 12 August and 4 September 2026, were included in all analyses.

Officers were recruited through internal emails. Thirteen were men and five were women, and their median age was 46 years (interquartile range 40 to 48; age was not available for three officers). Most had served for more than ten years (21 years or more, 8 officers; 11 to 20 years, 7; 6 to 10 years, 2; 3 to 5 years, 1). Five had received training on speech, language and communication needs from a speech and language therapist or Registered Intermediary, and 12 had worked on a case with a Registered Intermediary. All participants gave written informed consent, and the study was approved by [committee name] ([reference]).

The Forensic Interview Trainer

FIT runs on the research team's own hardware, without cloud services: a server with a single consumer graphics card (NVIDIA RTX 5070 Ti) runs all of the models, and a Raspberry Pi 5 client at the interview location captures the officer's speech and connects to the server over an encrypted virtual private network, so that no audio or text passed through third-party or cloud services. The officer's speech was transcribed with Whisper (19), and the transcript was passed both to the auditor described below and to a large language model, Qwen3.2 9B (q4), served locally through llama.cpp, which generated Jordan's reply. The reply was spoken by a zero-shot voice-cloning model, F5-TTS (20), and lip-synchronised with Wav2Lip (21) onto a still portrait of Jordan generated with SDXL (22), and the resulting video was played to the officer on a monitor (Figs. 1–3). Each reply took several seconds to produce, which limited the number of questions an officer could ask in five minutes.

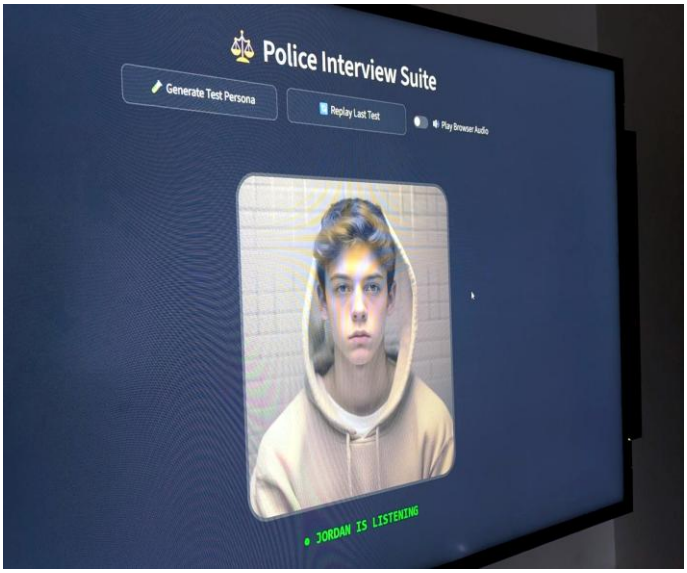


Fig. 1. The officer's view of FIT: Jordan's reply, spoken by a synthesised voice and lip-synchronised to a generated portrait, as shown on the interview display.



Fig. 2. The interview client: a Raspberry Pi 5 with an audio capture board and a USB push-button panel, which records each question and sends it to the server running the language model.



Fig. 3. A member of the research team demonstrating the interview set-up.

Jordan was specified as a 19-year-old man with undiagnosed DLD who had been arrested outside a bar late in the evening on suspicion of sexual assault, after a young woman alleged that he had put his hand on her waist and tried to kiss her. He gives short, fragmented answers of no more than eight to ten words and does not volunteer that he has failed to understand a question. Regional dialect was deliberately left out of the persona. During development, instructing the model through its prompt alone produced poor adherence to the persona's rules, with compliance of around 45% for some rules, so Jordan's replies were made hybrid: the model generates a candidate reply, and deterministic code then checks the officer's question with the auditor and shortens or replaces the reply to give the response the persona requires for each flagged feature. For example, the persona specifies that Jordan expresses confusion when asked a negatively phrased question, and agrees without engaging with the content when asked a tag question. The rules were refined against a fixed battery of 70 test questions derived from the clinical specification of the persona, on which they achieved over 95% compliance, so the pilot was the first test of Jordan's behaviour against unscripted questioning.

The auditor is a rule-based program that scores each question for the five features that best-practice guidance advises against: closed questions, questions with a high number of information-carrying words (ICWs), negatively phrased questions, tag questions and difficult vocabulary. It classifies a question as open if its first word is a question word (for example “what”, “how” or “why”) or the phrase “tell me”, and as closed otherwise. It reports an ICW count, but this counts multi-part questions rather than ICWs in the speech and language therapy sense: a question scores one or more only if it contains a second request introduced by “and” or “or” followed by a clause. It detects negation and tag questions from fixed word lists, and it flags as difficult any word with a Zipf frequency below 4.4 in the wordfreq database (23), after reducing the word to its base form. A question with any of the last four features is classified Not Recommended, an open question with none of them Recommended, and a closed question with none of them Partially Recommended. Before the pilot, the auditor agreed with all 19 of the persona specification’s worked examples. The auditor’s flags drove the enforcement layer during both interviews, its scores were shown to officers after the second interview, and, after the corrections described below, it supplied the study’s outcome measures.

Procedure

Each officer took part individually in a single session at The School of Electronics, Electrical Engineering and Computer Science (EEECS, Computer Science Building, 16A Malone Road, Belfast, Northern Ireland, BT9 5BN). Before the first interview, the officer was given [the scenario brief: what they were told about Jordan and the allegation, and in what form] and asked to interview Jordan as they normally would. The baseline interview then ran for five minutes, during which the officer saw only Jordan’s video replies and none of the auditor’s output. The officer then viewed a short presentation of slides, which described Jordan’s profile and the effects of DLD on comprehension and set out the five strategies corresponding to the auditor’s features, illustrated with examples such as opening questions with “Tell me”, “What” or “How” and replacing “witness” with “saw”. The officer then conducted a second five-minute interview with Jordan, about the same allegation, in a new FIT session. Only after the second interview did the researcher show the officer the Session Scoring view of the FIT Session Viewer for both interviews. This gave the number of questions the auditor had classified Recommended, Partially Recommended and Not Recommended, and for each question its ICW count, negation and tag flags, any words flagged as difficult and its classification, all as scored by the original auditor. The post-session coaching report that FIT can generate was not shown. Finally, the officer completed a questionnaire developed for this study (Additional file 1). It covered length of service and prior training; confidence in recognising and adapting to speech, language and communication needs, rated in retrospect for before the training and again for after it; agreement with statements about the training; whether they would recommend it in its current form; and open-ended comments.

Data preparation

FIT logs each turn of an interview as a separate record containing the transcript of the officer’s question, Jordan’s reply and the auditor’s results, and the 1,130 turns logged across all recorded sessions were cleaned with four rules before analysis. First, turns with no transcribed speech, and turns whose transcript was a transcription artefact or speech not addressed to Jordan, such as the sound of the interview timer, were removed (49 turns). Second, because a FIT session remained open until the researcher started the next one, microphone checks and other set-up speech between participants had been recorded into the previous participant’s session, so each interview was taken as its longest continuous run of turns with no gap of more than five minutes, and turns outside it were removed (15 turns). Third, greetings that formed the whole of an interview’s opening turns, such as “Hello Jordan”, were removed because they are not questions (18 turns), while greetings combined with a question were kept. Fourth, the sessions of the three non-police volunteers were excluded (121 questions). The resulting dataset contained 927 questions from 18 officers. Recording of one officer’s second interview stopped after 2.8 minutes, apparently because of a technical fault. The 13 questions recorded were retained, and a sensitivity analysis excluded this officer.

Auditor check and corrections

Because the auditor both drove the simulated suspect and supplied the outcome measures, it was checked before any comparison between interviews, using questions pooled across both timepoints. Re-running the original auditor on the

927 analysed questions reproduced every classification stored during the pilot, which confirmed that a single version of the auditor had scored all the interviews. Inspecting the flagged questions and words revealed two systematic errors, and review by a co-author identified a third. The first arose because the auditor judged a question to be open or closed from its first word alone, so that questions opening with a filler or the suspect's name, such as "Okay, what happened next?" or "Jordan, tell me about Saturday", were scored as closed, and a leading "and" was read as introducing a second question. The second arose because word frequency did not separate difficult words from everyday ones: the auditor flagged names, places and words central to the scenario, such as "nightclub", along with transcription errors, while 45% of its vocabulary flags fell on words that van Heuven et al. (24) class as high frequency (Zipf 4 or above). The third was that prompts beginning "explain" or "describe", which with "tell" make up the TED prompts used to invite free recall in investigative interviewing [citation], were scored as closed.

Three corrections were applied. First, leading fillers (for example "okay", "so", "right" and "Jordan") were removed before each question was scored, unless the turn consisted of fillers alone. Second, a flagged word was counted as difficult only if it is typically learned late, using the age-of-acquisition ratings of Kuperman, Stadthagen-Gonzalez and Brysbaert (25), obtained through the LexOPS dataset (26). Age of acquisition was chosen because the vocabulary of people with DLD is persistently narrower and shallower than that of their peers (27), so the words most likely to be unknown are those acquired latest. The threshold was set at 8.74 years, the rating of "witness", which the briefing slides gave as an example of a word to avoid, so that the cut-off followed the training's own definition rather than being chosen from the data. Because some ratings did not reflect British usage, six everyday British words rated as late-acquired (pub, pubs, bloke, lager, nightclub and nightclubs) were not counted as difficult, and words with no rating, mainly names and transcription errors, were not counted either. Third, questions opening with "tell", "explain" or "describe" were counted as open. The auditor's original decision rule was then re-applied to the corrected features. The corrected scores were used for all main analyses, and two sensitivity analyses repeated them with the original auditor's scores and with a stricter age-of-acquisition threshold of 10 years.

Statistical analysis

The primary outcome was the proportion of each officer's questions classified Recommended in each interview. For the first research question, the within-officer change in this proportion from baseline to post was summarised as a mean difference with a bias-corrected and accelerated bootstrap 95% confidence interval (10,000 resamples) (28) and as a standardised effect size for paired data (d_z), with a Wilcoxon signed-rank test as a check that did not depend on the distribution of the differences. Because questions are nested within officers and officers asked different numbers of questions, the same comparison was also made at the question level with a logistic model fitted by generalised estimating equations (29), with timepoint as the predictor, an exchangeable working correlation within officers, and the bias-reduced standard errors of Mancl and DeRouen (30), which perform better with few clusters; this is reported as an odds ratio with a 95% confidence interval. For the second research question, the same paired and question-level analyses were applied to the proportion of questions that were closed, contained a high number of ICWs as detected by the auditor (multi-part questions) or contained difficult vocabulary, with a Holm adjustment across these three features (31). Negated and tag questions were too rare for inference and are reported descriptively, and because the features differed in how common they were at baseline, comparisons of improvement between features are descriptive. Questions per interview are also reported. Because the auditor's ICW count did not measure information load directly, words per question and content words per question (nouns, proper nouns, main verbs, adjectives, adverbs and numbers, identified with spaCy [citation]) were also compared between interviews, the latter as an approximate indicator of the number of ICWs. All analyses were repeated with the original auditor's scores and with the stricter vocabulary threshold, and the primary analysis was also repeated without the officer whose second interview was only partly recorded. With 18 pairs and a two-sided α of 0.05, the study had 80% power only for within-officer effects of about $d_z = 0.70$ or larger, and the analyses are interpreted as estimates rather than as tests of effectiveness.

Five further sets of analyses were added after the planned analyses had been run, and they are reported as exploratory. The first described Jordan's replies, coded from their text for length, minimal replies of two words or fewer, "don't know" replies, expressions of confusion and bare yes or no answers, and compared them between questions with and without each feature, with timepoints pooled. The second examined whether officers changed behaviours that the

original auditor had penalised in error, such as opening a question with a filler or using a word wrongly flagged as difficult. The third tested whether question quality changed within the baseline interview, before the briefing, using a logistic model of question quality by minute of the interview. The fourth examined what officers asked immediately after Jordan failed to answer, a non-answer being a reply in which Jordan said he did not know, for example “dunno” or “no idea”, or signalled that he had not understood. The classification of the next question was modelled on whether the previous reply was a non-answer, adjusting for the previous question’s classification, with the same model specification as above; the proportions of negated questions, and of questions checking the suspect’s understanding or vulnerability (identified by keywords such as “understand”, “diagnosis” and “appropriate adult”), were compared after a non-answer and otherwise; and the question pairs were read to describe the strategies officers used. The fifth summarised the questionnaire. Confidence ratings before and after the training were compared with Wilcoxon signed-rank tests, treating “somewhat confident” and “moderately confident”, which occupied the same position on the before and after scales, as equivalent. Questionnaire answers and participant characteristics were related to officers’ baseline performance and improvement using Spearman correlations, with percentile bootstrap confidence intervals, and Mann–Whitney tests with rank-biserial correlations, without adjustment for multiple comparisons. Officers who reported that their scores had been affected by speech recognition errors were identified by reading their open-ended comments. Analyses were run in Python [3.12.8] with pandas, SciPy, statsmodels and spaCy in marimo notebooks.

Results

Interviews

The 18 officers asked 927 questions in total, 438 at baseline and 489 after training (Table 1). Interviews lasted a median of 4.8 minutes from first to last question at both timepoints, and some ran slightly over five minutes because officers were allowed to finish the question in progress. After training, officers asked slightly more questions (median 28 per interview, range 13 to 36, against 25, range 13 to 32, at baseline), and their questions were shorter (median of each officer’s mean words per question 7.2, range 4.7 to 9.5, against 10.6, range 7.9 to 33.8).

Measure	Baseline	Post
Questions, total	438	489
Questions per interview, median (range)	25 (13–32)	28 (13–36)
Mean words per question, median (range)	10.6 (7.9–33.8)	7.2 (4.7–9.5)
Interview length in minutes, median (range)	4.8 (3.9–7.0)	4.8 (2.8–5.3)

Table 1. Questions and interview length at each timepoint (18 officers). Interview length is measured from the first to the last question.

Auditor corrections

The three corrections changed the classification of 207 of the 927 questions (22%), and every change was upwards. Applied on its own, the filler correction changed 97 questions, the vocabulary correction 100 and the TED correction 7, all of them after training, and some questions needed more than one correction. Under the corrected rules, 344 questions (37.1%) were classified Recommended, 477 (51.5%) Partially Recommended and 106 (11.4%) Not Recommended, compared with 221 (23.8%), 477 (51.5%) and 229 (24.7%) under the original auditor (Table 2), and with the stricter vocabulary threshold of 10 years the corresponding figures were 352, 491 and 84. Corrections were more frequent at baseline (116 of 438 questions, 26.5%) than after training (91 of 489, 18.6%). The original auditor therefore under-rated the officers’ questioning at both timepoints, but more so at baseline: of all questions, it classified 14.6% at baseline and 32.1% after training as Recommended, against 29.0% and 44.4% under the corrected rules. These original scores were the ones officers were shown after their second interview.

Classification	Original auditor	Corrected (8.74 years)	Corrected (10 years)
Recommended	221 (23.8%)	344 (37.1%)	352 (38.0%)
Partially Recommended	477 (51.5%)	477 (51.5%)	491 (53.0%)
Not Recommended	229 (24.7%)	106 (11.4%)	84 (9.1%)

Table 2. Classification of the 927 questions under the original auditor and under the corrected auditor at two age-of-acquisition thresholds, both timepoints pooled. Values are n (%).

Overall change in questioning (research question 1)

The proportion of each officer's questions classified Recommended rose from a mean of 28.2% at baseline to 46.1% after training, a mean within-officer increase of 18.0 percentage points (95% CI 8.1 to 31.0; d_z = 0.71; Wilcoxon p = 0.004), and 15 of the 18 officers improved (Fig. 4). The question-level model gave a consistent estimate (odds ratio 2.03, 95% CI 1.36 to 3.03; p = 0.001). Of all questions, the share classified Not Recommended fell from 17.4% to 6.1%, while the share classified Partially Recommended changed little (53.7% to 49.5%). The increase was of similar size with the stricter vocabulary threshold (18.3 percentage points, 95% CI 9.1 to 31.6) and with the original auditor's scores (18.6 percentage points, 95% CI 10.5 to 25.8), under which 16 officers improved. Without the officer whose second interview was only partly recorded, the increase was smaller but still clear (13.6 percentage points, 95% CI 3.7 to 20.4; 14 of 17 officers improved).

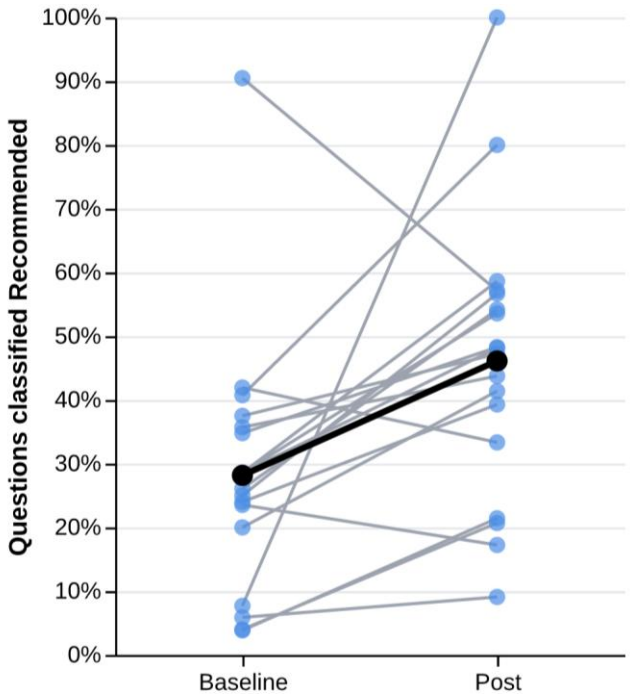


Fig. 4. Proportion of each officer's questions classified Recommended by the corrected auditor at baseline and after training. Grey lines join each officer's two interviews (n = 18), and the black line shows the mean.

Change in individual strategies (research question 2)

All three tested features became less common after training, and the question-level models agreed (Table 3, Fig. 5). In absolute terms the largest change was in closed questions, which fell from a mean of 69.5% of each officer's questions to 52.0% (−17.5 percentage points, 95% CI −31.1 to −8.4; d_z = −0.72; Holm-adjusted p = 0.010), although closed

questions remained the majority after training. Questions with difficult vocabulary fell from 11.2% to 3.8% (−7.4 percentage points, 95% CI −11.5 to −2.6; Holm-adjusted p = 0.010), and multi-part (high-ICW) questions from 4.2% to 1.0% (−3.2 percentage points, 95% CI −6.5 to −1.1; Holm-adjusted p = 0.028), although the auditor detected few such questions at either timepoint (17 and 5). Negated questions (21 and 10) and tag questions (5 and none) were too rare to test, but both became less frequent. All three tested features also decreased under the original auditor's flags and with the stricter vocabulary threshold. The decreases were larger under the original flags for closed questions (−20.3 percentage points, 95% CI −28.2 to −11.6) and difficult vocabulary (−12.5, 95% CI −18.9 to −5.9), because the original auditor had counted filler-led questions and TED prompts as closed and everyday words as difficult. Measures of information load that did not depend on the auditor also fell. The mean number of words per question fell from 12.7 to 7.2 (−5.5, 95% CI −9.9 to −3.7; d_z = −0.92; p < 0.001), and the mean number of content words, an approximate indicator of ICWs, fell from 4.7 to 2.7 (−2.1, 95% CI −3.8 to −1.2; d_z = −0.80; p < 0.001).

Feature	Baseline, n (mean %)	Post, n (mean %)	Change, pp (95% CI)	d_z	Holm p (Wilcoxon)	OR (95% CI)	Holm p (GEE)
Closed question	301 (69.5)	263 (52.0)	-17.5 (-31.1 to -8.4)	-0.72	0.010	0.51 (0.35 to 0.74)	< 0.001
Multi-part (high-ICW) question	17 (4.2)	5 (1.0)	-3.2 (-6.5 to -1.1)	-0.56	0.028	0.25 (0.07 to 0.91)	0.035
Difficult vocabulary	45 (11.2)	17 (3.8)	-7.4 (-11.5 to -2.6)	-0.73	0.010	0.32 (0.16 to 0.62)	0.002
Negation	21 (5.5)	10 (2.1)	-3.5	–	–	–	–
Tag question	5 (1.1)	0 (0.0)	-1.1	–	–	–	–

Table 3. Questions with each of the five features at baseline and after training (18 officers). n is the number of questions with the feature, and mean % is the mean across officers of the percentage of each officer's questions with it. Change is the mean within-officer difference in percentage points (pp), post minus baseline, with a bias-corrected bootstrap 95% CI. Holm-adjusted p values are adjusted across the three tested features. OR is the odds ratio for post against baseline from the question-level model. Negation and tag questions were too rare to test.

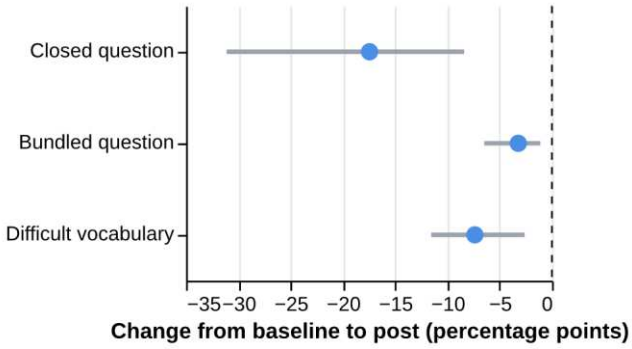


Fig. 5. Mean within-officer change in the percentage of questions with each tested feature from baseline to after training, with bias-corrected bootstrap 95% confidence intervals (n = 18). Negative values mean officers used the feature less. The dashed line marks no change.

Jordan’s responses (exploratory)

Jordan’s replies differed by question type in the way the persona intended for some features but not others (timepoints pooled). Closed questions drew shorter replies than open questions (mean 2.4 against 4.3 words), more minimal replies of two words or fewer (50% against 17%) and all but one of the bare yes or no answers, and all five tag questions drew a bare yes or no. Questions with difficult vocabulary and negated questions were answered with "don't know" almost without exception (100% and 97%), and multi-part (high-ICW) questions more often than other questions (59% against 23%). Replies this uniform reflect the enforcement layer rather than the language model, and negated questions drew “don’t know” rather than the expression of confusion the persona specifies. Because the enforcement layer acted on the original auditor’s flags, its errors reached the officers during the interview. Of the 170 questions it flagged for vocabulary, 108 contained no difficult word under the corrected rule, being flagged for names, everyday words such as “nightclub”, “pub” and “hoodie”, or transcription errors, and Jordan answered 99% of these with "don't know", against 7% of questions with no vocabulary flag. As a result, 12% of all the officers' questions received an automatic refusal caused by an auditor error, including questions that mentioned the nightclub at the centre of the scenario. Transcription errors also reached Jordan directly, since he responded to the transcript rather than to what the officer said, so that “High Street”, transcribed as “homestreet”, drew “don’t know”. No reply was coded as an expression of confusion.

Behaviours the original auditor penalised in error (exploratory)

The largest change in the dataset was in a behaviour the briefing never mentioned. The mean proportion of an officer’s questions that opened with a filler such as “okay”, “so” or “Jordan” fell from 51.1% at baseline to 17.6% after training (mean change −33.4 percentage points, 95% CI −44.3 to −21.6; Wilcoxon $p < 0.001$). Opening with a filler was the behaviour that the original auditor had most often penalised in error, by scoring such questions as closed (Table 4). Officers also used fewer words that the original auditor wrongly flagged as difficult (14.8% to 9.7% of questions; −5.1 percentage points, 95% CI −8.9 to −1.0; $p = 0.030$) and mentioned the nightclub less often (3.9% to 0.6%; −3.3, 95% CI −6.7 to −1.1; $p = 0.028$), while the reduction in questions under-rated by the original auditor was smaller and uncertain (26.4% to 19.4%; −6.9, 95% CI −16.5 to 3.2; $p = 0.12$). Despite these changes, the improvement measured by the original auditor was almost the same as that measured by the corrected auditor (18.6 against 18.0 percentage points; difference 0.6, 95% CI −10.1 to 10.9).

Behaviour	Baseline, mean %	Post, mean %	Change, pp (95% CI)	Wilcoxon p
Starts with a filler (okay, so, Jordan…)	51.1	17.6	-33.4 (-44.3 to -21.6)	< 0.001
Under-rated by the original auditor	26.4	19.4	-6.9 (-16.5 to 3.2)	0.119
Uses a word wrongly flagged as difficult	14.8	9.7	-5.1 (-8.9 to -1.0)	0.030
Mentions the nightclub	3.9	0.6	-3.3 (-6.7 to -1.1)	0.028

Table 4. Behaviours penalised in error by the original auditor, at baseline and after training (18 officers). Values are the mean across officers of the percentage of each officer’s questions showing the behaviour. Change is the mean within-officer difference in percentage points (pp), post minus baseline, with a bias-corrected bootstrap 95% CI and a Wilcoxon signed-rank test. These analyses are exploratory and unadjusted for multiple comparisons.

Change within the baseline interview (exploratory)

There was no clear evidence that officers improved during the baseline interview, before the briefing. The proportion of Recommended questions in the second half of the baseline interview was 7.5 percentage points higher than in the first (95% CI −3.6 to 16.2; $p = 0.12$). The question-level model gave an odds ratio of 1.19 per minute (95% CI 1.01 to 1.41), but this trend came from the final minute, which contained questions from a few officers whose interviews overran. In the second interview, by contrast, questions were already more often Recommended in the first minute (about 52%)

than in any minute of the baseline interview (about 24% to 37%), with no further change between the two halves of the interview (0.9 percentage points, 95% CI −10.6 to 12.8).

Officers' reactions to non-answers (exploratory)

Jordan gave a non-answer to 31% of questions at baseline and 18% after training. At baseline, a question that followed a non-answer was less likely to be Recommended than one that did not (odds ratio 0.42, 95% CI 0.21 to 0.81, adjusting for the previous question's classification), and this pattern did not change reliably after training (interaction odds ratio 1.66, 95% CI 0.72 to 3.82). Reading the pairs of questions showed four ways officers reacted:

- narrowing to a closed yes/no question, which was the most common;
- challenging Jordan, often with a negated question such as "Why do you not remember…?";
- re-explaining the allegation using legal terms;
- checking for a vulnerability, as in "Do you have any diagnosis of any communication issues?".

Simplifying the question successfully was less common. At baseline, negated questions made up 11.1% of questions after a non-answer (14 of 126) against 2.4% otherwise, and after training 5.8% (5 of 86) against 1.3%. Vulnerability checks made up 14.3% of questions after a non-answer against 6.1% otherwise at baseline, but showed no difference after training (2.3% against 2.6%), when officers had been told that Jordan has DLD. The auditor scores such checks as closed questions.

Officers' views of the training (exploratory)

Officers rated the training very positively (Table 5). All 18 agreed or strongly agreed that it was interactive, that it gave them direct feedback, that the feedback was useful, that it reflected challenges they could encounter in policing, and that it could benefit less experienced colleagues, and 15 agreed that it gave them new strategies for recognising and adapting to speech, language and communication needs. Asked whether they would be comfortable recommending the tool in its current form, 11 officers answered yes or mainly yes, four were not sure and three answered mainly no. All seven who did not answer yes gave technical reasons: speech recognition errors, Jordan's failure to follow on from earlier answers, difficulty with the regional accent, or a need for refinement before operational use. Self-rated confidence in recognising speech, language and communication needs without a diagnosis was higher after the training for 8 officers, unchanged for 9 and lower for 1 (Wilcoxon p = 0.020), and confidence in adapting questioning was higher for 6, unchanged for 11 and lower for 1 (p = 0.058). The median rating was "quite confident" before and after.

The training…	Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree	Agree or strongly agree, n (%)
Gave me new strategies for recognising SLCN	0	0	3	10	5	15 (83%)
Gave me new strategies for adapting my communication	0	0	3	10	5	15 (83%)

Made me aware of challenges I might not usually notice	0	0	1	13	4	17 (94%)
Made me reflect on the clarity of my own communication	0	0	1	10	7	17 (94%)
Was interactive	0	0	0	11	7	18 (100%)
Gave me an opportunity to practise adapting my communication	0	0	1	10	7	17 (94%)
Gave me direct feedback about my communication	0	0	0	10	8	18 (100%)
The feedback was useful in helping me adapt	0	0	0	12	6	18 (100%)
Reflected challenges I could encounter in policing practice	0	0	0	6	12	18 (100%)
Could be a useful tool for learning about SLCN	0	0	0	5	13	18 (100%)
Could benefit less experienced colleagues	0	0	0	3	15	18 (100%)
I would feel comfortable using it in future	0	0	0	7	11	18 (100%)

My colleagues would feel comfortable using it in future	0	1	0	10	7	17 (94%)

Table 5. Officers' agreement with statements about the training (n = 18). Values are numbers of officers. SLCN, speech, language and communication needs.

Officers' comments corroborated the scoring errors identified in the auditor check. Without being asked, six officers wrote that questions had been marked down because their words had been misheard ("Some of my not recommended questions were actually misheard", officer 18), and these six did have more of their questions under-rated by the original auditor (median 27.7% against 20.6%; rank-biserial r = 0.64; p = 0.032). Six officers also wrote that, faced with Jordan in a real interview, they would not continue without an assessment, a Registered Intermediary or an appropriate adult. Officers' characteristics were largely unrelated to their questioning (Additional file 2). Neither years of service (ρ = 0.02 with baseline performance, ρ = 0.18 with improvement), age, gender nor experience of working with a Registered Intermediary showed a clear relationship. The five officers who had received training from a speech and language therapist or Registered Intermediary asked more Recommended questions at baseline than the other 13 (median 37.5% against 24.0%; r = 0.75; p = 0.018), but did not improve more (median change 9.7 against 17.6 percentage points; p = 0.70). Self-rated confidence in adapting questioning was only weakly related to baseline performance (ρ = 0.28, p = 0.26), changes in confidence were only weakly related to measured improvement (ρ = 0.22 and 0.24), and whether an officer would recommend the tool was unrelated to their improvement (ρ = 0.03).

Discussion

In this exploratory pilot, a short briefing on questioning people with communication needs, given between two five-minute interviews with a language model-simulated suspect with DLD, was associated with a large improvement in officers' use of questions suited to such a suspect. The proportion of Recommended questions rose by 18.0 percentage points, and 15 of the 18 officers improved, although the estimate was lower, at 13.6 percentage points, without the one officer whose second interview was only partly recorded. Officers also asked shorter questions with fewer content words, in keeping with the briefing's advice to keep questions short and simple. The improvement was present from the first minute of the second interview rather than building up during it, and there was no clear improvement during the first interview, which suggests that the briefing, rather than practice with Jordan, was its main source, although the design cannot separate the two. Earlier simulation studies attributed improvements in questioning to feedback given during training (14,15). Here, officers received no individual feedback until after their second interview, so the gain followed a single practice interview and a short briefing. Unlike Røed et al (15), who found that open questions increased while closed questions did not fall, closed questions fell in the present study, although they still made up more than half of officers' questions after training. Whether so quick a gain lasts is unknown, since interviewing practice declines once feedback and supervision stop (13). Because the design had no control group, and the outcome measured exactly what the briefing taught, the size of the improvement should be read as an upper bound on what the training achieved, not as an estimate of its effect on real interviews.

The most consequential finding concerns the automated scoring. The auditor misclassified more than one question in five, always under-rating the officer, and because its flags also drove the enforcement layer, its errors shaped the simulation itself: 12% of officers' questions received an automatic "don't know" caused by an everyday word, a name or a transcription error, including questions about the nightclub at the centre of the scenario. Officers had no way of telling these refusals apart from genuine difficulties of comprehension, and in their second interview they used fewer of the words wrongly flagged as difficult and mentioned the nightclub less often, before they had seen any scores, which suggests that the simulation was, in part, teaching them to avoid words that a suspect with DLD would understand. The errors had less effect on the measured improvement, which was almost the same under the original and corrected scoring, than on the level of performance and on the scores officers were finally shown, which under-rated their

questioning and penalised questions that had simply been misheard. Officers noticed: six said so unprompted, and those six had indeed been under-rated more often, yet all 18 agreed that the feedback was useful. Automated feedback is therefore likely to be trusted even when trainees can see that it is partly wrong, and it should be validated against expert judgement before it is used to teach, as Røed et al.(15) did by checking their question classifier against human coders ($\kappa = 0.80$) before using it. The same applies to the measures themselves: the auditor's ICW count detected only multi-part questions, and although the fall in content words suggests that the information load of officers' questions did decrease, counting ICWs properly requires clinical judgement of the conversational context.

The training also changed the surface form of officers' questions in a way that the scoring happened to reward. The largest change in the study was a fall in questions opening with a filler such as "okay" or the suspect's name, a behaviour the briefing never mentioned and that officers could not have learned from scores they had not yet seen. The most likely explanation is that officers copied the question openings illustrated on the briefing slides, all of which began with the question word, and that the auditor's first-word rule then rewarded the new form. The briefing taught the same five features that the auditor measured, so the outcome measured exactly what was taught, which is the situation in which a measure risks becoming a target (32). Dropping "okay" or the suspect's name may also strip out the small markers of rapport on which suspect interviewing depends (33), and a question can begin with "what" and still be long, leading or hard to follow, so improvement on the auditor's terms should not be equated with better interviewing.

The simulation also exposed how officers reacted when the suspect failed to answer. After a non-answer, officers most often narrowed to a closed yes/no question, or challenged the suspect with a negated question, or re-explained the allegation in legal terms. These are the responses that best-practice guidance advises against for people with communication needs (8), and they resemble the reliance on closed questions seen in recorded police interviews with suspects (12). Successful simplification was less common. Officers also checked for a vulnerability more often after a non-answer at baseline. This is arguably good practice, yet the auditor scored it as a closed question, which illustrates how a rubric built around question form can penalise behaviour that the underlying guidance would welcome. Six officers went further, writing that in a real interview they would not have continued with Jordan without an assessment, a Registered Intermediary or an appropriate adult. Stopping to secure communication support may be the right response to a suspect like Jordan, yet the simulation offered no way to do so and the auditor gave no credit for it. These reactions were not reliably changed by the training, and training that addresses how to repair a breakdown in communication, and when to stop and seek support, rather than only the form of individual questions, may be a more useful target for future versions.

The study cannot say how this form of training compares with conventional police training, since there was no comparison arm. Simulation does offer things that classroom instruction does not. It gives repeated practice with a standardised suspect, an immediate consequence when a question is poorly constructed, and feedback on every question asked, and it could provide the ongoing feedback that sustained good practice appears to require (13) at a cost that does not depend on an expert's time. Officers valued these features, describing the training as a safe way to practise without role players, and all agreed that the interaction reflected challenges they could encounter in policing, although several missed the tone, facial expression and non-verbal cues that real interviewees provide. These advantages hold only if the simulated suspect demonstrates the speech, language and communication needs associated with DLD and the feedback is accurate, and the present results show that neither can be assumed. The officers who had received specialist training from a speech and language therapist or Registered Intermediary questioned better from the outset, whereas length of service made no difference, which suggests that simulation may complement such training rather than replace it. Officers' confidence was also a poor guide to their performance, so evaluations of interview training that rely on officers' self-reports may misjudge its effects.

The study has further limitations:

- **Sample:** 18 officers from a single force, most with long service. Their characteristics, including prior training, were self-reported.
- **Scope:** each officer interviewed one suspect in one scenario, for five minutes at a time, and with no follow-up, so retention is unknown. One officer's second interview was only partly recorded.

- **Feedback:** officers saw their scores only after the second interview, so the study did not test whether automated feedback itself improves questioning.
- **The corrected auditor:** it still has known weaknesses. It detects a high number of ICWs only in multi-part questions joined by “and” or “or”, and misses even these when they are separated only by punctuation or contain contractions, and it treats a question that opens with a “when” clause as open.
- **Vocabulary norms:** the age-of-acquisition ratings were not collected from British raters, and some did not reflect British usage.
- **Jordan’s realism:** this was not assessed against the communication of people with DLD, and his replies to negated questions departed from the persona’s specification, giving “don’t know” rather than an expression of confusion. His speech was synthetic and stilted, and each reply took several seconds.
- **Transcription:** errors were passed to both Jordan and the auditor.
- **Questionnaire:** it was developed for this study and has not been validated. All its statements were worded positively, officers completed it with the research team present, and the “before” confidence ratings were given in retrospect.
- **Coding:** non-answers and vulnerability checks were identified by keywords rather than by human coders.
- **Exploratory analyses:** these were decided after the data were seen, and several rest on small numbers of questions or officers.

Future work should compare FIT with an active control, and test whether automated feedback given between interviews adds to the effect of the briefing. It should validate the auditor against coding by speech and language therapists, have Jordan’s transcripts rated for realism by speech and language therapists and by people with lived experience of DLD, drive the enforcement layer from a validated scorer, and allow officers to pause the interview and request communication support.

Conclusions

A short briefing given between two interviews with a language model-simulated suspect with DLD was associated with a large improvement in police officers’ use of questions suited to suspects with communication needs, and officers valued the training highly. However, errors in the automated scoring shaped both the simulated suspect’s replies and the scores officers were shown, and officers trusted that feedback even when they could see that it was partly wrong. The uncontrolled design cannot isolate the contribution of the simulation, and improvement on the auditor’s terms may partly reflect a change in the surface form of questions rather than in their quality. The simulation also revealed how officers instinctively respond when a suspect does not answer, and that several would stop and seek communication support, both potential targets for training. AI-mediated interview training should validate its automated scoring against expert judgement, and test the realism of its simulated interviewees, before its feedback is used to teach.

Abbreviations

AI: artificial intelligence; CI: confidence interval; DLD: developmental language disorder; FIT: Forensic Interview Trainer; ICW: information-carrying word; OR: odds ratio; pp: percentage points; SLCN: speech, language and communication needs; TED: tell, explain, describe

Declarations

Ethics approval and consent to participate

The study was approved by [committee name] ([reference]) and conducted in accordance with the Declaration of Helsinki. All participants gave written informed consent before taking part.

Consent for publication

Written consent to publish the photograph in Fig. 3 was obtained from the member of the research team shown. The image of Jordan in Figs. 1 and 2 was generated by computer and does not depict a real person. Participants consented to the publication of anonymised quotations [confirm against the consent form].

Availability of data and materials

The interview transcripts and questionnaire responses are not publicly available because they could identify participants, who sometimes stated their names during interviews and belong to a small group of officers from a single force. Aggregate results and the analysis code, written as marimo notebooks, are available via private GitLab repo on request. The age-of-acquisition norms were obtained from the LexOPS dataset, released under a CC BY 4.0 licence.

Competing interests

CB and FS developed the Forensic Interview Trainer evaluated in this study. The system is not commercially available, and the authors declare no other competing interests.

Funding

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Authors' contributions

CB conceived the study, developed the FIT software, carried out the analysis and drafted the manuscript. FS conceived the study, wrote the clinical specification of the persona, collected the data and revised the manuscript. KR and JT supervised the work and revised the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We thank the officers who took part, and the three volunteers who took part in the technical dry runs.

References

1. Bishop DVM, Snowling MJ, Thompson PA, Greenhalgh T, Consortium and the C 2. Phase 2 of CATALISE: a multinational and multidisciplinary Delphi consensus study of problems with language development: Terminology. J Child Psychol Psychiatry. 2017;58(10):1068–80. doi:10.1111/jcpp.12721

2. Norbury CF, Gooch D, Wray C, Baird G, Charman T, Simonoff E, et al. The impact of nonverbal ability on prevalence and clinical presentation of language disorder: evidence from a population study. J Child Psychol Psychiatry. 2016;57(11):1247–57. doi:10.1111/jcpp.12573

3. Anderson SAS, Hawes DJ, Snow PC. Language impairments among youth offenders: A systematic review. Child Youth Serv Rev. 2016 Jun 1;65:195–203. doi:10.1016/j.childyouth.2016.04.004

4. Bryan K, Freer J, Furlong C. Language and communication difficulties in juvenile offenders. Int J Lang Commun Disord. 2007;42(5):505–20. doi:10.1080/13682820601053977

5. Lavigne M, Rybroek GJV. Breakdown in the Language Zone: The Prevalence of Language Impairments Among Juvenile and Adult Offenders and Why It Matters.

6. Winstanley M, Webb RT, Conti-Ramsden G. More or less likely to offend? Young adults with a history of identified developmental language disorders. Int J Lang Commun Disord. 2018 Mar;53(2):256–70. doi:10.1111/1460-6984.12339 PubMed PMID: 29159847; PubMed Central PMCID: PMC5888152.

7. Clare ICH, Gudjonsson GH, Harari PM. Understanding of the current police caution (England and Wales). J Community Appl Soc Psychol. 1998;8(5):323–9. doi:10.1002/(SICI)1099-1298(1998090)8:5<323::AID-CASP448>3.0.CO;2-2

8. Toolkits | The Advocate's Gateway [Internet]. [cited 2026 Oct 17]. Available from: https://www.theadvocatesgateway.org/toolkits

9. Perry NW, McAuliff BD, Tam P, Claycomb L, Dostal C, Flanagan C. When lawyers question children: Is justice served? Law Hum Behav. 1995;19(6):609–29. doi:10.1007/BF01499377

10. Kebbell MR, Johnson SD. Lawyers' questioning: the effect of confusing questions on witness confidence and accuracy. Law Hum Behav. 2000 Dec;24(6):629–41. doi:10.1023/a:1005548102819 PubMed PMID: 11105476.

11. Northern Ireland Registered Intermediary Scheme | Department of Justice [Internet]. 2016 [cited 2026 Oct 17]. Available from: https://www.justice-ni.gov.uk/articles/northern-ireland-registered-intermediary-scheme

12. Snook B, Luther K, Quinlan H, Milne R. Let 'em talk!: A field study of police questioning practices of suspects and accused persons. Crim Justice Behav. 2012;39(10):1328–39. doi:10.1177/0093854812449216

13. Lamb ME, Sternberg KJ, Orbach Y, Esplin PW, Mitchell S. Is ongoing feedback necessary to maintain the quality of investigative interviews with allegedly abused children? Appl Dev Sci. 2002;6(1):35–41. doi:10.1207/S1532480XADS0601_04

14. Pompedda F, Zappalà A, Santtila P. Simulations of child sexual abuse interviews using avatars paired with feedback improves interview quality. Psychol Crime Law. 2015;21(1):28–52. doi:10.1080/1068316X.2014.915323

15. Røed RK, Baugerud GA, Hassan SZ, Sabet SS, Salehi P, Powell MB, et al. Enhancing questioning skills through child avatar chatbot training with feedback. Front Psychol. 2023 Jul 13;14. doi:10.3389/fpsyg.2023.1198235

16. Holderried F, Stegemann-Philipps C, Herschbach L, Moldt JA, Nevins A, Griewatz J, et al. A Generative Pretrained Transformer (GPT)-Powered Chatbot as a Simulated Patient to Practice History Taking: Prospective, Mixed Methods Study. JMIR Med Educ. 2024 Jan 16;10:e53961. doi:10.2196/53961 PubMed PMID: 38227363; PubMed Central PMCID: PMC10828948.

17. Shanahan M, McDonell K, Reynolds L. Role play with large language models. Nature. 2023 Nov;623(7987):493–8. doi:10.1038/s41586-023-06647-8 PubMed PMID: 37938776.

18. Zeng J, Qi W, Shen S, Liu X, Li S, Wang B, et al. Embracing the Future of Medical Education With Large Language Model-Based Virtual Patients: Scoping Review. J Med Internet Res. 2025 Nov 13;27:e79091. doi:10.2196/79091 PubMed PMID: 41232097; PubMed Central PMCID: PMC12661241.

19. Radford A, Kim JW, Xu T, Brockman G, McLeavey C, Sutskever I. Robust Speech Recognition via Large-Scale Weak Supervision.

20. Chen Y, Niu Z, Ma Z, Deng K, Wang C, Zhao J, et al. F5-TTS: A Fairytaler that Fakes Fluent and Faithful Speech with Flow Matching [Internet]. arXiv; 2025 [cited 2026 Oct 29]. Available from: http://arxiv.org/abs/2410.06885 doi:10.48550/arXiv.2410.06885

21. Prajwal KR, Mukhopadhyay R, Namboodiri V, Jawahar CV. A Lip Sync Expert Is All You Need for Speech to Lip Generation In The Wild. In: Proceedings of the 28th ACM International Conference on Multimedia [Internet]. 2020 [cited 2026 Oct 29]. p. 484–92. Available from: http://arxiv.org/abs/2008.10010 doi:10.1145/3394171.3413532

22. Podell D, English Z, Lacey K, Blattmann A, Dockhorn T, Müller J, et al. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis [Internet]. arXiv; 2023 [cited 2026 Oct 29]. Available from: http://arxiv.org/abs/2307.01952 doi:10.48550/arXiv.2307.01952

23. Speer R. rspeer/wordfreq: v3.0 [Internet]. 2022 Sep 26. doi:10.5281/zenodo.7199437

24. van Heuven WJB, Mandera P, Keuleers E, Brysbaert M. SUBTLEX-UK: a new and improved word frequency database for British English. Q J Exp Psychol. 2014;67(6):1176–90. doi:10.1080/17470218.2013.850521 PubMed PMID: 24417251.

25. Kuperman V, Stadthagen-Gonzalez H, Brysbaert M. Age-of-acquisition ratings for 30,000 English words. Behav Res Methods. 2012 Dec 1;44(4):978–90. doi:10.3758/s13428-012-0210-4

26. Taylor JE, Beith A, Sereno SC. LexOPS: An R package and user interface for the controlled generation of word stimuli. Behav Res Methods. 2020 Dec;52(6):2372–82. doi:10.3758/s13428-020-01389-1 PubMed PMID: 32394182; PubMed Central PMCID: PMC7725696.

27. McGregor KK, Oleson J, Bahnsen A, Duff D. Children with developmental language impairment have vocabulary deficits characterized by limited breadth and depth. Int J Lang Commun Disord. 2013;48(3):307–19. doi:10.1111/1460-6984.12008 PubMed PMID: 23650887; PubMed Central PMCID: PMC3648877.

28. Efron B. Better Bootstrap Confidence Intervals. J Am Stat Assoc. 1987;82(397):171–85. doi:10.2307/2289144

29. LIANG KY, ZEGER SL. Longitudinal data analysis using generalized linear models. Biometrika. 1986 Apr 1;73(1):13–22. doi:10.1093/biomet/73.1.13

30. Mancl LA, DeRouen TA. A covariance estimator for GEE with improved small-sample properties. Biometrics. 2001 Mar;57(1):126–34. doi:10.1111/j.0006-341x.2001.00126.x PubMed PMID: 11252587.

31. Holm S. A Simple Sequentially Rejective Multiple Test Procedure. Scand J Stat. 1979;6(2):65–70.

32. Strathern M. 'Improving ratings': audit in the British University system. Eur Rev. 1997 Jul;5(3):305–21. doi:10.1002/(SICI)1234-981X(199707)5:3<305::AID-EURO184>3.0.CO;2-4

33. Alison L, Alison E, Noone G, Elntib S, Waring S, Christiansen P. The efficacy of rapport-based techniques for minimizing counter-interrogation tactics amongst a field sample of terrorists. Psychol Public Policy Law. 2014;20(4):421–30. doi:10.1037/law0000021