Developing Annotation Guidelines for CSAM Prevention Interventions: Psychosocial Risk and Protective Factors Grounded in Research and Clinical Practice

Vera Czehmann[1,3], Christine Hovhannisyan[4], Lena Hoffmann[3], Paula Busch[2], Ibrahim Baroud[1,3], Sebastian Möller[1,3], Roland Roller[1], Hannes Gieseler[2], Lisa Raithel[1,3,5,6]

[1]German Research Center for Artificial Intelligence (DFKI GmbH),
[2]Charité – Universitätsmedizin Berlin, Institute of Sexology and Sexual Medicine,
[3]Quality & Usability Lab, Technische Universität Berlin, [4]Humboldt-Universität zu Berlin,
[5]BIFOLD – Berlin Institute for the Foundations of Learning and Data,
[6]Charité – Institut für Künstliche Intelligenz in der Medizin (IKIM)
Berlin, Germany
{vera.czehmann, roland.roller}@dfki.de
christine.hovhannisyan@student.hu-berlin.de
l.hoffmann@campus.tu-berlin.de
{paula.busch, hannes.gieseler}@charite.de
{ibrahim.baroud, raithel}@tu-berlin.de

Abstract

This work discusses sexual offending, specifically child sexual abuse material (CSAM), in the context of prevention. We introduce a domain-specific, span-level annotation scheme and guidelines to identify psychosocial *risk* and *protective* factors in therapist-led, anonymous chat interventions with voluntarily help-seeking individuals concerned about their pedophilic interests and the risk of CSAM use. The scheme is grounded in previous research and clinical experience, and intended for within-intervention guidance and longitudinal tracking, rather than actuarial risk scoring. Annotating a pilot subset (8 clients, 31 sessions), inter-annotator agreement was moderate but improved after calibration, which is consistent with the linguistic and clinical ambivalence present in the data. We track a session-wise *Protective Ratio*, i.e., the share of protective factors among all coded factors, and examine its behaviour over time during the intervention and around self-reported relapse within clients. In exploratory automation, LLM-based span extraction outperforms BERT baselines but overall performance remains limited by small data and mixed-evidence spans. While complete anonymisation of the corpus is in progress, we release the label scheme, guidelines, and non-sensitive artefacts of our analyses.

Keywords: CSAM prevention, psychosocial risk and protective factors, annotation guidelines, span extraction, grounded methodology, therapist expertise

1. Introduction and Motivation

Sexual interest in minors (commonly described as pedophilic or hebephilic interest) constitutes a persistent sexual preference pattern rather than, in itself, a criminal act (Jahnke, 2018). Such interests are classified under paraphilic disorders only when the person has acted on their urges, they cause distress, impairment, or involve risk of harm to themselves or others (American Psychiatric Association, 2022). Acting on such urges can manifest in different forms, including contact child sexual abuse (CSA) and the use of child sexual abuse material (CSAM); images or other media depicting the abuse or exploitation of minors. While these behaviours are interconnected, research recognises distinctions between contact and non-contact offending populations in terms of risk profiles, motivations, and intervention needs (Babchishin et al., 2015).

A growing body of research also distinguishes between forensic populations, who enter the system following detected CSAM-related offences, and clinical populations. Within this so-called *Dunkelfeld* (the "dark figure" of undetected offences), there exists a subset of individuals actively seeking help to prevent CSAM-related (re)offending (Von Franqué et al., 2023). In our clinical context, we refer to this as *relapse*, i.e., a return to CSAM use, as self-reported by clients. Many such individuals experience fear of disclosure (Jahnke, 2018). In response, several prevention-oriented programmes have emerged to offer confidential and, in some cases in their online extensions, anonymous therapeutic or self-guided interventions for individuals seeking to manage their attraction responsibly. These initiatives reveal an underserved population of help-seeking individuals whose communications with counsellors provide a unique window into cognitive, emotional and behavioural cues relevant for relapse prevention. Understanding which factors could predict positive or negative trajectories is critical for effective intervention.

Proceedings of the 6th RaPID@MENTAI.ai Workshop @ LREC 2026, pages 126–145
12 May 2026.

We differentiate between risk factors that could increase the likelihood of relapse, and protective factors, resources, strategies, or cognitions that could help prevent or reduce the likelihood of relapse. This work explores how such factors are expressed linguistically in preventive CSAM intervention chats and whether they can be tied to intervention outcome and self-reported relapse behaviours. We introduce a domain-specific annotation scheme tailored to these dialogues and report inter-annotator agreement and statistical analyses on detected factors.

The scheme is intended to support the manual creation of a gold standard set that could train and evaluate automatic classifiers to produce a large-scale, transparent corpus. Prospectively, this data could be utilised in building a tool to help flag risk and protective factors in client talk, support therapists' decision-making during sessions and enable the longitudinal tracking of client progress with respect to therapist intervention. We furthermore provide a *first baseline using Large Language Models for extracting the defined factors*. Finally, we conducted a *semi-structured expert interview*, as previously done by Klymenko et al. (2022). References to the interview or expert opinion will be marked throughout with an asterisk (*). The full annotation scheme and guidelines and a transcript of the interview can be found on Zenodo[1].

2. Related Work

Unlike other medical domains with routinised outcomes and large public datasets, preventive therapeutic interventions regarding sexual offences lack shareable corpora, despite increasing interest in dynamic risk modelling and online interventions. Therefore, this work bridges three strands of prior work: (i) client language and outcomes, (ii) forensic risk-assessment constructs, and (iii) secondary prevention in the *Dunkelfeld*.

Client language and intervention outcomes. A recent survey of mental health datasets emphasises that while labelled datasets on multi-class classification and questionnaire score prediction exist, there is a particular scarcity in genuine therapy corpora (Mandal et al., 2025). Research on online, text-based counselling shows that client language can be annotated reliably and is predictive of subsequent outcomes. For Motivational Interviewing (MI), a therapy strategy used, e.g., in the treatment of addiction, Wu et al. (2023) released a dataset of transcribed counselling dialogue demonstrations, expert-annotated for MI-specific concepts such as change- and sustain-centered client talk on the dialogue and utterance levels. Previous works in the field of MI have found that while an association with change talk has not been consistently reported across studies, sustain talk was positively associated with worse outcome (Magill et al., 2018).

Ewbank et al. (2021) manually coded transcripts of internet-enabled Cognitive Behavioural Therapy (CBT) for five categories of client utterances, informed by the MI technique (Amrhein et al., 2003). This was then utilised in training a deep-learning classifier to auto-code transcripts at scale. Model performance reached human-level agreement on most of the categories and, crucially, the automatically derived signals were reliably linked to outcomes. They also identified demographic predictors of reliable improvement from the first session. Together, these findings indicate that span-level labels on client talk are both feasible and informative for downstream support tools.

Risk assessment in the forensic field. Forensic risk-assessment frameworks used with convicted sexual offenders offer constructs relevant to, among others, sexual preoccupation, cognitive distortions, and self-regulation. Actuarial tools (e.g., STATIC-99R) quantify risk from static, file-based information (Phenix et al., 2017). Structured professional judgement (SPJ) approaches (SVR-20; RSVP) use standardised item sets in combination with decision guidance to support clinician risk formulation (Hart and Boer, 2020). Dynamic instruments (STABLE-2007; ACUTE-2007) capture relatively stable and acute changeable factors, and recent work links their scores to recidivism among men adjudicated for CSAM offences (Babchishin et al., 2023). For CSAM-specific recidivism, the Child Pornography Offender Risk Tool (CPORT) operationalises offence-relevant items and shows predictive utility (Seto and Eke, 2015). Recent review work has also synthesised psychosocial characteristics and risk-related profiles in detected CSAM offenders (Barroso et al., 2026). These tools and findings conceptually inform our annotation scheme. However, they are not directly transferable to anonymous, therapist-led prevention chats with undetected offenders or individuals at risk of CSAM-related offending (Von Franqué et al., 2023).

Secondary prevention in the *Dunkelfeld*. Secondary prevention services reach voluntarily help-seeking individuals outside the justice system. Evaluations report decreases in offence-supporting cognitions and emotional deficits, alongside gains in self-regulation, although a residual risk of CSAM use (relapse) may persist (Beier et al., 2015; Von Franqué et al., 2023). Therapist-led, chat-based anonymous interventions broaden access, with CBT approaches showing initial reductions in

[1] https://zenodo.org/records/19189153

CSAM viewing among motivated users (Lätth et al., 2022). Other studies indicate substantial demand and characterise user profiles, including factors linked to help-seeking and motivation to stop (Insoll et al., 2024), as well as higher distress and CSAM use disclosures among users with pedophilic or hebephilic interests (Schuler et al., 2021).

Prototype therapist-assistive AI tools for chat-based interventions in the domain reduce clinicians' perceived cognitive load (Deshpande et al., 2025a), and retrieval-augmented LLM suggestions can match or exceed therapist replies (Deshpande et al., 2025b). These findings underscore that therapist support at scale requires machine-actionable, span-level labels that index risk or protective factors.

3. Data

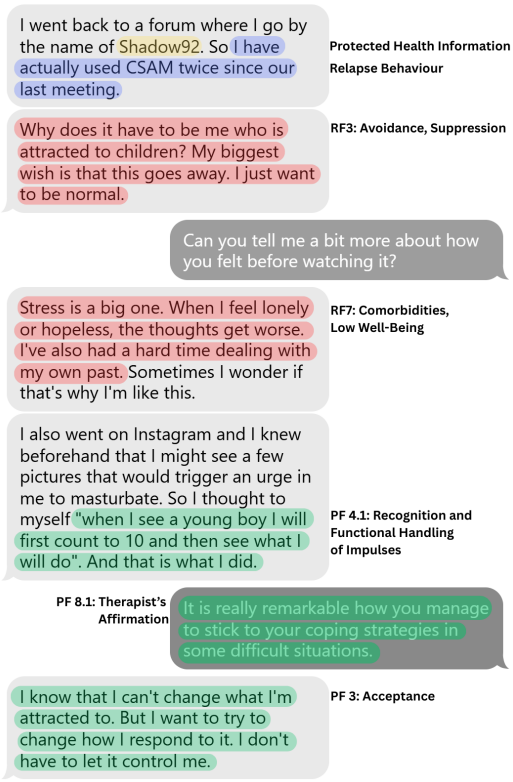

Figure 1: A fictional example excerpt of a session between a therapist and a client with annotated risk and protective factors.

We tailored our annotation scheme to chat transcripts from intervention chats between help-seeking individuals and therapists. They were collected within an anonymous online study on preventive support for individuals self-reported to be at risk of CSAM-related offending, conducted by the Institute of Sexology and Sexual Medicine at Charité – Universitätsmedizin Berlin. A fictional example of one such therapy chat is shown in Figure 1.

Clients	Sessions	Tokens	Tokens per Session
8	31	62,034	2,001

Table 1: Corpus size and typical session length.

Participants enrol voluntarily and have the right to withdraw at any time. Eligible individuals provide informed consent to the use of their chat transcripts for research. Over the course of a 12 week period, participants receive access to self-help material and a varying amount of scheduled chat sessions with a therapist of 50 minutes each. Both therapists and clients remain anonymous throughout. Chat logs are stored on secure institutional servers without personal identifiers and are automatically de-identified, replacing detected PHI with placeholders before using them for annotation. We retain anonymised speaker and time metadata to capture turn-taking.

Although a substantially larger, multilingual corpus is being collected, for this first annotation study we focus on a randomly selected subset of eight English speaking clients (details in Table 1). While we cannot release raw transcripts at this time, we are conducting iterative de-identification to annotate and remove remaining Protected Health Information[2] as well as indirect identifiers (as suggested by Baroud et al., 2025) with the goal of sharing an extensively anonymised subset in the future, subject to ethics approval and data sharing agreements.

4. Guideline Development

We next describe the process of developing our annotation scheme and guidelines. Factors are intended as clinically interpretable anchors and actionable labels for session guidance and longitudinal tracking, not as actuarial or forensic risk scores.

4.1. Annotation Scheme and Guidelines

We target a preventive, non-forensic setting: voluntarily help-seeking individuals concerned about their (risk of) CSAM use. Our proposed annotation scheme is shown in Table 2. The first version was drafted by a team member with extensive clinical experience working with individuals at risk of CSAM-related offending in both offline and online settings. Combined with this, a targeted synthesis of prior research informed a first-pass taxonomy of linguistically observable risk and protective factors indicative of potential relapse behaviour, suitable for span-level annotation in therapist-led chats. In our annotation, we prioritised concepts that clinicians actively monitor during prevention work, and which can be expressed explicitly or implicitly in

[2]PHI, annotation guidelines by Lohr et al. (2024).

client talk and could be actionable for session-by-session tracking.

The complete annotation guidelines include further descriptions of factors, fictional example spans for each factor, and guides on prioritisation between some semantically adjacent factors. Definitions of some factors also include client talk of recognising the importance of certain protective concepts, even if it did not reflect in their actions. While the proposed annotation scheme is domain-specific and purpose-built for preventive chat interventions, all factors are conceptually grounded in prior research.

4.2. Grounding in Existing Literature

Sexual preference for children (RF1) has been found to be more prevalent in users of CSAM, with previous work reporting that online offenders score higher on measures of pedophilic interest (Schuler et al., 2021; Babchishin et al., 2015). Regarding **preference patterns** (RF1.1), the CPORT, having shown predictive utility for sexual recidivism among CSAM offenders, includes items on male-focused interest as a risk-enhancing factor (Seto and Eke, 2015). Consequently, **non-exclusivity**, present sexual interest in adults (PF1), is treated as a protective factor consistent with strengths-based rehabilitation (Ward et al., 2025; Willis and Ward, 2024) and emerging work on dynamic protective factors that reduce reliance on illegal material by opening pathways to viable, lawful intimacy (Thornton et al., 2024).

The Good Lives Model (Ward et al., 2025) frames capability-building for prosocial, consensual relationships as incompatible with offending. Thus, mentions of **healthy intimacy** (PF2) mark a protective trajectory. Clinic-centred prevention reports similarly target relational functioning and empathy as change mechanisms in voluntarily help-seeking populations (Von Franqué et al., 2023; Beier et al., 2015). **Difficulties forming or maintaining trusting, mutually supportive adult relationships** (RF2, 2.1) are treated in structured professional judgement (SPJ) and dynamic frameworks as relatively stable vulnerabilities (Babchishin et al., 2023; Hart and Boer, 2020).

Non-acceptance, active avoidance or suppression of the sexual preference in client statements (RF3) is treated as risk relevant, as it can undermine self-regulation and increase distress, making planned coping harder (Hart and Boer, 2020). Shame- and avoidance-driven presentations could also reduce engagement and follow-through, with clinical prevention reports suggesting that moving towards an **accepting, integrated self-image** (PF3) can support consistent coping and value-congruent behaviour (Von Franqué et al., 2023; Beier et al., 2015).

SPJ frameworks consider problems with stress or coping a risk factor in the psychological adjustment domain (Hart and Boer, 2020). Consistent with this, we annotate **dysfunctional coping and impulse control deficits** (RF4, 4.1), because relapses often occur when self-regulation fails in the presence of acute triggers*. CBT models describe the operational mechanism of high-risk situations paired with ineffective coping leading to relapse, whereas specific, **healthy coping** responses (PF4) can interrupt the chain (Marlatt and Donovan, 2005). In the CSAM context, early evidence from internet-delivered CBT shows that equipping motivated users with concrete coping strategies can reduce viewing of problematic content (Lätth et al., 2022).

Cognitive validation (RF5) captures offence-supportive attitudes and minimisations that reduce perceived wrongfulness or harm, which has been found to be an empirically supported risk factor for sexual recidivism (Mann et al., 2010). SPJ frameworks also consider denial of sexual violence and attitudes that condone violence as risk factors (Hart and Boer, 2020). Qualitative interviews suggest that there are implicit theories in child abusers that account for the majority of their cognitive distortions (Marziano et al., 2006). In CSAM specific work, reviews report closely related cognitions in online offending (Bartels and Merdian, 2016). We code **recognition of harm** (PF5) as protective counterpart.

Past problematic behaviour, specifically CSAM use, and **criminal history** (RF6) are consistently linked to higher recidivism risk in CSAM literature (Babchishin et al., 2015). The CPORT operationalises case file-derived predictors (e.g., offence history) and has demonstrated predictive utility for sexual recidivism among CSAM offenders (Helmus et al., 2025; Eke et al., 2019). We code as protective factors (PF6, 6.1) when clients articulate **healthy sexual behaviour** or **abstinence from illegal material explicitly for fear of legal consequences**, reflecting observations that such client talk may relate to more favourable trajectories*.

Comorbidities and low well-being (RF7) are annotated because co-occurring burdens (e.g., depressed mood, anxiety, or substance use) can weaken self-regulation and amplify triggers, increasing relapse risk*. This is consistent with instruments that flag negative affect and substance use (Babchishin et al., 2023) and with SPJ guidance integrating psychosocial adjustment and mental disorders into risk assessment (Hart and Boer, 2020). Complementarily, we annotate explicit statements of **well-being** (PF7) that go beyond courtesy phrases.

Poor therapy and change commitment, and externalisation (RF8, 8.1) captures general agency-distancing client talk. SPJ guidance treats

Risk Factors	
RF0	Not specified
RF1	Sexual preference for children
RF1.1	Sexual preference for male children
RF2	Intimacy deficits, lack of trustful social relationships
RF2.1	Being in a dysfunctional relationship
RF3	Avoidance, suppression, missing acceptance of the sexual preference
RF4	Dysfunctional coping (strategies)
RF4.1	Lack of impulse control
RF5	Reduction of cognitive dissonance, cognitive validation of problematic behaviour
RF6	Past problematic behaviour, criminal history
RF7	Comorbidities, low well-being, other psychological problems or diseases
RF8	Poor therapy and change commitment, missing confidence about reaching the goals and changing behaviour
RF8.1	Directing responsibility to someone else, externalising problems
RF9	Sociodemographic factors
RF10	Hypersexuality, sexual preoccupation
RF10.1	Failure to satisfy sexual needs in a healthy way
RF11	Hostility, preoccupation towards other groups
RF12	Compulsive sexualisation of non-sexual content (of children) and/or situations (with children)

Protective Factors	
PF1	Non-exclusivity of the sexual preference for children
PF2	Healthy intimacy, trustful social relationships and acknowledgement of the importance of it
PF3	Acceptance of the sexual preference
PF4	Functional coping (strategies)
PF4.1	Recognition and functional handling of impulses
PF5	Recognition of the abuse of children in the material
PF6	Healthy sexual behaviour
PF6.1	Abstinence of problematic behaviours due to fear of legal consequences
PF7	Well-being
PF8	Cooperation with therapist, commitment to the treatment or study setting
PF8.1	Therapist's affirmation of commitment
PF10	Skills to satisfy sexual needs (urges) in a healthy way without harming themselves and/or others

Table 2: Overview of the proposed annotation scheme with risk (left) and protective (right) factors.

issues with treatment or supervision and negative attitudes towards the intervention as risk indicators (Hart and Boer, 2020). In MI research, sustain talk and therapist-client discord track poorer outcomes, whereas **commitment** (PF8) and reason or need language predict improvement (Miller, 2023; Magill et al., 2018). Analyses on internet-enabled CBT similarly show that motivated client talk relates to better outcomes (Ewbank et al., 2021). Specific to chat-delivered intervention, we also code **therapist's affirmation of commitment** (PF8.1).

We consider several risk-relevant **sociodemographic factors** (RF9). While the CPORT considers age of 35 or younger at time of the index investigation as an item related to higher risk of recidivism (Seto and Eke, 2015), we follow the observations of involved therapists that relatively young age at the time of the intervention could be associated with negative outcomes*. Informed by SPJ guidance, we also annotate problems with employment (Hart and Boer, 2020), and housing.

In risk factor syntheses, **hypersexuality** (RF10) is identified as a correlate (Mann et al., 2010), and CSAM-focused reviews note how sexual preoccupation, in combination with availability and anonymity online, sustains use and can co-occur with escalation of the unwanted behaviour in severity or frequency (RF10.1) (Helmus et al., 2025; Baskurt et al., 2025). Correspondingly, we explicitly code mentions of **skills to satisfy sexual needs in a healthy way** without harming themselves or others (PF10), countering preoccupation by enabling concrete, value-consistent choices (Lätth et al., 2022).

Hostility-laden preoccupation (RF11), the combination of hostile affect and sexual focus, is theoretically well-grounded. The confluence model links hostile masculinity and impersonal or **sexualised cognition** to elevated risk for sexual aggression (Malamuth et al., 1996). We code this as compulsive sexualisation of non-sexual content or situations (RF12).

5. Annotation

The annotation was carried out by a team of five trained annotators. The team included both domain experts and trained research assistants, ensuring that clinical expertise and methodological

rigor were combined throughout the annotation process. One of the annotators was a physician and experienced therapist actively involved in the chat interventions from which the data were drawn, providing first-hand contextual insight into the therapeutic setting. The remaining annotators comprised a research associate in clinical psychology, a final year Bachelor's student in psychology, a final year Master's student in computer science, and a computer science student with a medical background.

All annotators were fluent in English and received detailed instruction on the annotation guidelines prior to beginning the task. Before the main annotation phase, they participated in a structured training phase that included guideline familiarisation, collective discussion of example cases, and pilot annotations on a small subset of the data. This pilot phase served to refine both the guidelines and the annotators' shared understanding of the categories to be applied.

Annotation was performed using *INCEpTION*[3], which supports span-level annotations with type (RF or PF) and feature assignments (exact labels of respective factors). All data was stored on an institute-owned server with restricted use and access only via the *INCEpTION* interface. Regular calibration meetings were held throughout the annotation process to discuss ambiguous cases, ensure consistency, and update the guidelines.

The dataset was divided into multiple subsets for annotation. Two annotators independently annotated one subset[4], and two different annotators independently annotated a second subset. A small portion of the data was annotated by all four annotators who took part in the second iteration to facilitate comprehensive reliability assessment. The task of the annotators was to (i) identify relevant spans in the conversations between therapist and client, (ii) decide whether the span was a protective or a risk factor and, finally, (iii) decide which factor exactly was represented, following the definitions in Table 2 and further hints from our annotation guidelines. Descriptions of CSAM use since the last session were annotated separately from the factor scheme as *Relapse Behaviour*. Additionally, two annotators labelled the entire session with a binary document-level label for relapse (*known relapse* or *no relapse* since the last session).

Across the 31 annotated sessions, annotators identified 1,670 factor instances in total, comprising 775 risk factors and 895 protective factors. This corresponds to an overall pooled protective ratio of 0.54, indicating a slight predominance of protective over risk-related client talk in the pilot corpus. We report the current status and results of annotations and corresponding challenges after two iterations of group discussion and guidelines as well as scheme refinement in the following. Further results and considerations are reported in Appendix A.

Inter-annotator agreement (IAA) was calculated on overlapping subsets of the data to monitor annotation reliability and to guide iterative revisions of the scheme. For this, we used *Krippendorff's* α *(unitizing)* (Krippendorff et al., 2016) computed natively in *INCEpTION*, giving credit for partial overlap. To analyse IAA in more detail, we also computed *relaxed* F_1[5] (Hripcsak and Rothschild, 2005). We performed label-aware 1:1 greedy maximum-overlap matching (>0 character overlap) in three modes (*general*, *exact*, *categorised*). True positives were counted as overlap and agreement on *exact* factor label, *general* type (RF/PF), or *category*. We aggregated results as pooled micro-F_1, and session-macro. Differences across iterations and between *exact* vs. *categorised* were quantified with bootstrap percentile 95% CI.

	i_1	i_2	$\Delta(i_2 - i_1)$
$\langle pos \rangle$	0.33 [0.22–0.41]	0.60 [0.53–0.67]	0.27*** [0.15–0.40]
RF	0.15 [0.02–0.26]	0.36 [0.20–0.54]	0.22† [0.02–0.42]
PF	0.33 [0.25–0.41]	0.40 [0.27–0.54]	0.07 (n.s.) [-0.09–0.23]

Table 3: α_u. Improvement over iterations (i) per annotation dimension: $\langle pos \rangle$ for position, *RF* and *PF* for risk factor and protective factor, respectively. Means; 95% CI in brackets. Significance of Δ tested via permutation tests: $^{***}p < .001$, $^{\dagger}p < .10$, n.s. $=$ not significant.

Krippendorff's α *(unitizing)* agreement on the span-level *position* of annotations increased significantly across iterations (details in Table 3). Agreement on *risk factors* improved descriptively, approaching conventional significance. For *protective factors*, no robust improvement was observed. On one session annotated by four annotators within the second iteration, α_u scores reached 0.48 for position, 0.47 on risk factors and 0.40 on protective factors, respectively.

Session-macro micro- and macro-F_1 improved significantly across sessions at both the *general* (RF vs. PF) and the *exact* factor level (see Table 4). These gains indicate that calibration meetings held between iterations improved both the coarse decision between risk or protective factor and the finer-grained code assignments, with the largest absolute improvements observed at the *general* level.

[3] https://inception-project.github.io/; version 35.2.

[4] One subset was annotated by three annotators. Normalisation was performed when computing IAA.

[5] Used libraries are reported in Appendix A.

	i_1	i_2	$\Delta(i_2 - i_1)$
micro $_{general}$	0.56 [0.47–0.64]	0.68 [0.60–0.75]	0.13*** [0.05–0.20]
micro $_{exact}$	0.41 [0.32–0.49]	0.53 [0.45–0.61]	0.12** [0.03–0.21]
macro $_{general}$	0.52 [0.43–0.60]	0.67 [0.58–0.75]	0.15*** [0.05–0.23]
macro $_{exact}$	0.31 [0.21–0.40]	0.43 [0.33–0.53]	0.12* [0.01–0.23]

Table 4: Relaxed F_1. Improvement over iterations (i): on the type level (*general*, RF vs. PF) and the *exact* factor level. The rows labelled *micro* and *macro* report session-macro micro-F_1 and session-macro macro-F_1, respectively. Point estimates; 95% CI in brackets. Δ: session-level bootstraps; significance tested with unpaired bootstrap: $^{***}p < .001$, $^{**}p < .01$, $^{*}p < .05$.

Annotator confusion between factors and contrastive disagreements. In a subsequent effort to improve stability and annotator consistency, we tested grouping semantically similar fine-grained labels into higher-level categories within their type (RF and PF, respectively). Factors where no adjacent label was sufficiently similar to justify merging were retained as singleton categories. Pooling sessions from both iterations, agreement improved only in a modest, albeit consistent, manner.

In the second iteration of annotations, the main confusion hubs were risk factors of the category *Affect-/stress-driven coping and impulsivity*. Factors within *Self-regulation and functional coping* mirrored this on the protective side. Importantly, the factor with the most observed contrastive disagreements between risk and protective factors was *Recognition and functional handling of impulses* (PF), with relatively high counts of *confusions* with both *Dysfunctional coping (strategies)* (RF) and *Lack of impulse control* (RF).

6. Results

We present quantitative analyses of annotated risk and protective factors, tracking their distribution across sessions and around relapse, and report pilot span-extraction results with large language models (LLMs). Further results are reported in Appendix B.

6.1. *Protective Ratio* Over Time

To examine changes in the relative distribution of protective versus risk factors across sessions, we computed a *Protective Ratio (PR)* per client and session, $n_{PF}/(n_{PF} + n_{RF})$, where n_{PF} and n_{RF} are the numbers of annotated protective and risk factors.

At the session level, the *Protective Ratio* had a mean of 0.55, a median of 0.52, and ranged from 0.12 to 1.00.

A linear mixed-effects model revealed a significant positive effect of session number ($\beta = 0.11$, $SE = 0.03$, $p < .001$, 95% CI [0.05, 0.16]). This indicates that the relative proportion of protective factors systematically increased over time across all clients.

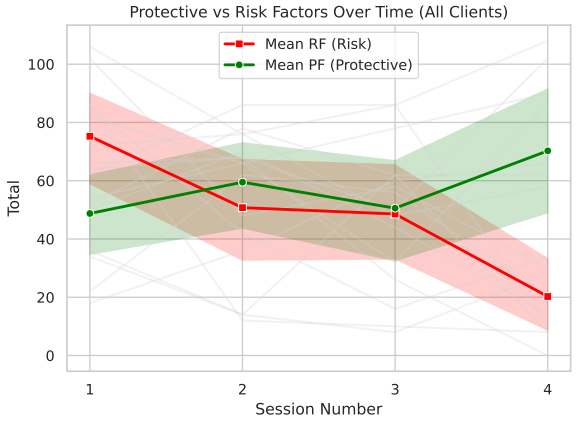

Figure 2: The trends of risk (red) and protective (green) factors over time, i.e., intervention sessions.

We also plotted mean per-client counts of protective and risk factors by session number to visualise their trajectories (Figure 2). In addition, we calculated monotonic trend tests for each individual client. While some clients showed clear monotonic increases in *PR*, others displayed more variable trajectories.

6.2. *Protective Ratio* Around Relapse

To examine whether the balance between protective and risk factors changes around relapse, we aligned sessions to self-reported relapses (pre- or post-relapse session windows) and contrasted them against all other sessions from the same client (baseline) using two-sided permutation tests. Differences (Δ) are computed as window minus baseline; negative values indicate a decrease from baseline. For the *Protective Ratio*, we summarise both a pooled, and the mean session-wise *PR*. In the pre-relapse window, both pooled and mean *PR* decreased relative to baseline (Δ = -0.10). In the post-relapse window, pooled *PR* was likewise lower (Δ = -0.07) and the mean *PR* showed the largest drop (Δ = -0.16). Though these effects are exploratory at the present pilot size, sessions immediately surrounding relapse contain a relatively higher share of risk than protective factors.

6.3. Factor-Level Analyses Around Relapse

For analyses on the *exact* factor level, we summarise the largest directional movers as exploratory effect sizes. We quantified relapse-aligned changes for each exact factor using two lenses: *presence* (difference in the proportion of sessions in which a factor appears), and normalised *rate* (difference in factor counts divided by total annotations in the respective session), indicating the relative dominance of these factors just before or after relapse.

For pre-relapse sessions, across lenses, the strongest changes were increases for mentions of *Avoidance, suppression, missing acceptance of the sexual preference* (RF3; $\Delta_{presence}$ = 0.52, Δ_{rate} = 0.05) and *Comorbidities, low well-being* (RF7; $\Delta_{presence}$ = 0.41, Δ_{rate} = 0.05), and smaller decreases across several protective factors. Post-relapse, *Cooperation and commitment* (PF8), *Skills to satisfy sexual needs in a healthy way* (PF10), *Recognition of the abuse of children* (PF5), and *Healthy sexual behaviour* (PF6) appeared less often, whereas *Dysfunctional coping* (RF4) and *Avoidance and non-acceptance of the sexual preference* (RF3) were comparatively more dominant. An across-client sensitivity analysis (global label shuffling) yielded similar rankings.

6.4. LLM-based Span Extraction

To explore the feasibility of automatic span extraction for risk and protective factors in CSAM-related therapeutic chats, we conducted pilot experiments using expert-annotated sessions as gold standard data.

To evaluate large language models (LLMs), we used *Qwen2.5:14b*[6] and *Mistral:7b*[7] in few-shot and fine-tuned setups. Prompts were enriched with examples and definitions derived from our annotation guidelines. *Qwen2.5:14b* served as the primary model, with *Mistral:7b* as a secondary baseline. Classification was performed both at the span level (*span*), and for entire messages (*message*), with separate runs for risk and protective factors (best results reported in Table 5).

In addition, we experimented with three BERT-based models. While these models could capture some patterns, their performance was substantially lower than that of the LLMs. Overall, protective factors were more accurately detected than risk factors. Further details, prompts, and results of the experiments will be reported in Appendix C.

[6] last accessed on 04.10.2025 via https://ollama.com/library/qwen2.5

[7] last accessed on 04.10.2025 via https://ollama.com/library/mistral

Model	Type / Level	$\mathbf{F_1}$ Score
Qwen2.5:14b (ft)	PFs (span)	0.350
Qwen2.5:14b (fs)	RFs (message)	0.296

Table 5: Best results of LLM span extraction experiments. "ft" refers to fine-tuned models, while "fs" refers to few-shot models.

7. Discussion and Outlook

Improved agreement across two annotation iterations indicates that annotator calibration meetings and revised, specific guidelines translate into better convergence. While absolute agreement scores remain moderate, we consider this a good result for pilot annotation, considering the task's linguistic subtlety and the breadth of the taxonomy. Importantly, many observed *disagreements* appear to reflect genuine clinical ambivalence rather than noise*. Several factors are best understood on a continuum, moving toward risk or protection depending on how clients express themselves on the same topic*. A concrete example are relationships: *trustful, fulfilled intimacy* functions protectively, whereas *lack or poor quality of intimacy* could increase risk. Simply "being in a relationship" is not protective without evidence of quality*. Whether a segment is coded as risk or protective factor frequently hinges on span boundaries as well, and whether annotators actually privilege intention or recognising the importance of certain concepts over actual behaviour as per our guidelines. Consequently, most *contrastive disagreements* happen around *coping and impulse control*; conscious reflection of dysfunctional coping could itself be protective, yet is easy to misread as risk*. These factors were also the most disputed in calibration meetings*.

Complementary analyses suggest that the scheme captures clinically coherent dynamics*. The *Protective Ratio* shows a systematic increase across sessions of all clients, while relapse-aligned windows show lower PR (a higher share of risk factors) in sessions before and after relapse. Clinically, post-relapse sessions often focus on relapse processing, which raises the salience of risk-coded talk even as it serves a therapeutic goal*. The pre-relapse elevation of clients expressing *Avoidance or non-acceptance* (RF3) and mentioning *Comorbidities or low well-being* (RF7) is intriguing against heterogeneous client profiles*, and indicates a possibility to detect factors that generalise across clients as early warning signals, given a larger database. While exploratory at current power, these patterns are directionally consistent with clinical experience*.

In pilot experiments, LLMs outperformed BERT-based models on span extraction, but absolute scores remain modest. This is plausible given the

small corpus, fine-grained labels, and the prevalence of mixed-evidence spans, and it mirrors the ambiguity seen in human annotation. Protective factors were detected more accurately than risk factors, which is consistent with their lower conceptual complexity and the smaller number of class distinctions. This pattern is also visible in IAA results, especially in the first annotation iteration.

Next, we will prioritise corpus growth, guideline tightening, and annotator training. Concretely, we will use fictional examples to synthesise diverse client profiles and intervention sessions. We will run adjudication sprints, and update the guidelines accordingly. Resulting annotations will be used for further annotator calibration, with a focus on span boundaries and contrastive risk versus protective factor decisions. To lower annotator cognitive load and improve label consistency, we suggest developing a flow-based decision aid that supports the coding of ambiguous spans and delivers tie-break rules. Future work could also examine client profiles to test whether specific constellations of factors are associated with elevated relapse risk, and further analyse criterion validity of the factors against relapse behaviour and therapist-judged relapse risk.

8. Conclusion

This work translates clinically meaningful risk and protective factors into span-level labels for therapist-led, prevention-oriented chats with voluntarily help-seeking individuals concerned about their risk of CSAM use. We introduce an annotation scheme and guidelines, ran a two-iteration pilot with targeted calibration, and observed reliability gains. Factor dynamics were examined using within-client baselines aligned to self-reported relapse. The *Protective Ratio* rises generally across sessions, but dips around relapse. Thus, the scheme is sensitive to session progression and relapse proximity and, with further guideline refinements and annotator training, can be used to create an actionable dataset. Although more annotated training and evaluation data are required before robust models are feasible, LLMs hold promise for identifying clinically relevant language patterns in prevention-oriented CSAM interventions.

Acknowledgements

We gratefully acknowledge funding from the German Federal Ministry of Research, Technology and Space (BMFTR) through the project VERANDA (16KIS2046K) and through the grant BIFOLD26B.

Limitations

We acknowledge several limitations of the present study. The current corpus is relatively small, which constrains the statistical robustness of our observations and limits the generalisability of the findings. Accordingly, some of the reported tendencies should be read as preliminary rather than conclusive. While inter-annotator agreement (IAA) remains comparatively low, we view this as an informative result in its own right. It reflects the conceptual and linguistic difficulty of coding psychosocial risk and protective factors in therapeutic discourse, where genuine ambivalence and span-boundary judgements are common. Model performance remains modest, reflecting both the small data size and the challenging, context-dependent nature of the task.

The dataset itself cannot yet be shared, as full anonymisation and completion of the annotation process are still ongoing. This necessary restriction currently limits reproducibility and external validation. However, we open-source non-sensitive artefacts related to this study, such as annotation guidelines, the label scheme, and results of our analyses.

Our scheme includes factors adapted from tools developed for forensic contexts. They are clinically plausible in our prevention-oriented *Dunkelfeld* chats, but not directly transferable. Base rates, help-seeking and disclosure incentives, and our span-level, language-based coding differ from file- or clinician-rated risk assessment. We therefore read our results with caution, as useful clinical cues rather than actuarial indicators. Future work should further establish validity in prevention cohorts.

Finally, potential sampling biases must be considered: the available data involve exclusively male clients, which may not fully represent the entire population of individuals seeking help for CSAM-related concerns. Despite these limitations, the study establishes a valuable foundation for future research on the linguistic identification of psychosocial risk and protective factors in prevention-oriented interventions.

Ethical Considerations

Working with data related to child sexual abuse material and associated risk/protective factors involves significant ethical and societal responsibility. The individuals whose data inform this work are part of a highly vulnerable and stigmatised population. All data were collected and processed in strict compliance with ethical guidelines and applicable data protection regulations. Prior to any analysis, data were de-identified and will undergo full anonymisation before potential publication or sharing. No

personally identifying information is disclosed at any stage of the project.

Due to the sensitive nature of the material, particular attention was paid to ensuring psychological safety for all researchers and practitioners involved in data handling and annotation. Regular open discussion was offered to minimise exposure-related distress and to maintain a high standard of ethical awareness.

Despite these challenges, this work addresses an ethically crucial domain: prevention and early intervention. Developing tools that can help detect psychosocial risk and protective factors in anonymised therapeutic communication has the potential to support therapists in their decision-making, contribute to offender prevention, and ultimately enhance victim protection. The overarching goal of this research is not surveillance or control, but the empowerment and support of therapeutic professionals and the advancement of evidence-based prevention efforts.

Ethical Approval. The chat study was reviewed and approved by the institutional ethics board of Charité – Universitätsmedizin Berlin under approval number EA4/187/24, ensuring compliance with all relevant ethical and legal standards for research involving sensitive and high-risk populations.

9. Bibliographical References

American Psychiatric Association. 2022. *Diagnostic and Statistical Manual of Mental Disorders*, dsm-5-tr edition. American Psychiatric Association Publishing.

Paul C. Amrhein, William R. Miller, Carolina E. Yahne, Michael Palmer, and Laura Fulcher. 2003. Client commitment language during motivational interviewing predicts drug use outcomes. *Journal of Consulting and Clinical Psychology*, 71(5):862–878.

Kelly M. Babchishin, Ségolène Dibayula, Chiara McCulloch, R. Karl Hanson, and L. Maaike Helmus. 2023. ACUTE-2007 and STABLE-2007 predict recidivism for men adjudicated for child sexual exploitation material offending. *Law and Human Behavior*, 47(5):606–618.

Kelly M. Babchishin, R. Karl Hanson, and Heather VanZuylen. 2015. Online Child Pornography Offenders are Different: A Meta-analysis of the Characteristics of Online and Offline Sex Offenders Against Children. *Archives of Sexual Behavior*, 44(1):45–66.

Ibrahim Baroud, Lisa Raithel, Sebastian Möller, and Roland Roller. 2025. Beyond De-Identification: A Structured Approach for Defining and Detecting Indirect Identifiers in Medical Texts. In *Proceedings of the Sixth Workshop on Privacy in Natural Language Processing*, pages 75–85, Albuquerque, New Mexico. Association for Computational Linguistics.

Ricardo Barroso, Sofia Silva, Mariam Fishere, Julia Nentzl, Thuy Nguyen Vo, Carlos García Forero, Esperanza Luísa Gómez-Durán, Catarina Braz Ferreira, Hannes Gieseler, Viola Westfal, Berta Franch-Martínez, Lucie Krejčová, and Klaus M Beier. 2026. Child sexual abuse material (CSAM): a systematic review of risk profiles, risk factors, and typologies of users. *Sexual Medicine Reviews*, 14(1):qeaf081.

Ross M. Bartels and Hannah L. Merdian. 2016. The implicit theories of child sexual exploitation material users: An initial conceptualization. *Aggression and Violent Behavior*, 26:16–25.

Serra Baskurt, Kelly M. Babchishin, Gabriella Hilkes, and Michael C. Seto. 2025. A meta-analysis of recidivism rates among individuals who commit child sexual exploitation material (CSEM) offending. *Aggression and Violent Behavior*, 85:102080.

Klaus M. Beier, Dorit Grundmann, Laura F. Kuhle, Gerold Scherner, Anna Konrad, and Till Amelung. 2015. The German Dunkelfeld Project: A Pilot Study to Prevent Child Sexual Abuse and the Use of Child Abusive Images. *The Journal of Sexual Medicine*, 12(2):529–542.

Neha Deshpande, Mariam Fishere, and Stefan Hillmann. 2025a. The Development of an AI-Assistant to Therapists in a Chat-based Psychological Intervention: Gathering Users' First Impressions of the Experience. Cagliari, Italy.

Neha Deshpande, Stefan Hillmann, and Sebastian Möller. 2025b. Evaluating Large Language Models for Enhancing Live Chat Therapy: A Comparative Study with Psychotherapists. In *Proceedings of the 26th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 800–812, Avignon, France. Association for Computational Linguistics.

Angela W. Eke, L. Maaike Helmus, and Michael C. Seto. 2019. A Validation Study of the Child Pornography Offender Risk Tool (CPORT). *Sexual Abuse*, 31(4):456–476.

M. P. Ewbank, R. Cummins, V. Tablan, A. Catarino, S. Buchholz, and A. D. Blackwell. 2021. Understanding the relationship between patient lan-

guage and outcomes in internet-enabled cognitive behavioural therapy: A deep learning approach to automatic coding of session transcripts. *Psychotherapy Research*, 31(3):300–312.

Stephen D. Hart and Douglas P. Boer. 2020. Structured professional judgment guidelines for sexual violence risk assessment: the sexual violence risk-20 (SVR-20) versions 1 and 2 and risk for sexual violence protocol (RSVP). In *Handbook of violence risk assessment*, pages 322–358. Routledge.

L. Maaike Helmus, Angela W. Eke, and Michael C. Seto. 2025. What risk assessment tools can be used with men convicted of child sexual exploitation material offenses? Recommendations from a review of current research. *Law and Human Behavior*, 49(1):71–88.

George Hripcsak and Adam S. Rothschild. 2005. Agreement, the f-measure, and reliability in information retrieval. *Journal of the American Medical Informatics Association: JAMIA*, 12(3):296–298.

Tegan Insoll, Valeriia Soloveva, Eva Díaz Bethencourt, Anna Katariina Ovaska, Juha Nurmi, Arttu Paju, Mikko Aaltonen, and Nina Vaaranen-Valkonen. 2024. Factors Associated with Help-Seeking Among Online Child Sexual Abuse Material Offenders: Results of an Anonymous Survey on the Dark Web. *Journal of Online Trust and Safety*, 2(4).

Sara Jahnke. 2018. The Stigma of Pedophilia: Clinical and Forensic Implications. *European Psychologist*, 23(2):144–153.

Oleksandra Klymenko, Stephen Meisenbacher, and Florian Matthes. 2022. Differential Privacy in Natural Language Processing The Story So Far. In *Proceedings of the Fourth Workshop on Privacy in Natural Language Processing*, pages 1–11, Seattle, United States. Association for Computational Linguistics.

Klaus Krippendorff, Yann Mathet, Stéphane Bouvry, and Antoine Widlöcher. 2016. On the reliability of unitizing textual continua: Further developments. *Quality & Quantity*, 50(6):2347–2364.

Christina Lohr, Franz Matthies, Jakob Faller, Luise Modersohn, Andrea Riedel, Udo Hahn, Rebekka Kiser, Martin Boeker, and Frank Meineke. 2024. De-Identifying GRASCCO – A Pilot Study for the De-Identification of the German Medical Text Project (GeMTeX) Corpus. In Rainer Röhrig, Niels Grabe, Ursula Hertha Hübner, Klaus Jung, Ulrich Sax, Carsten Oliver Schmidt, Martin Sedlmayr, and Antonia Zapf, editors, *Studies in Health Technology and Informatics*. IOS Press.

Johanna Lätth, Valdemar Landgren, Allison McMahan, Charlotte Sparre, Julia Eriksson, Kinda Malki, Elin Söderquist, Katarina Görts Öberg, Alexander Rozental, Gerhard Andersson, Viktor Kaldo, Niklas Långström, and Christoffer Rahm. 2022. Effects of internet-delivered cognitive behavioral therapy on use of child sexual abuse material: A randomized placebo-controlled trial on the Darknet. *Internet Interventions*, 30:100590.

Molly Magill, Timothy R. Apodaca, Brian Borsari, Jacques Gaume, Ariel Hoadley, Rebecca E. F. Gordon, J. Scott Tonigan, and Theresa Moyers. 2018. A meta-analysis of motivational interviewing process: Technical, relational, and conditional process models of change. *Journal of Consulting and Clinical Psychology*, 86(2):140–157.

Neil M. Malamuth, Christopher L. Heavey, and Daniel Linz. 1996. The Confluence Model of Sexual Aggression: Combining Hostile Masculinity and Impersonal Sex. *Journal of Offender Rehabilitation*, 23(3-4):13–37.

Aishik Mandal, Prottay Kumar Adhikary, Hiba Arnaout, Iryna Gurevych, and Tanmoy Chakraborty. 2025. A Comprehensive Survey of Datasets for Clinical Mental Health AI Systems. Version Number: 2.

Ruth E. Mann, R. Karl Hanson, and David Thornton. 2010. Assessing Risk for Sexual Recidivism: Some Proposals on the Nature of Psychologically Meaningful Risk Factors. *Sexual Abuse*, 22(2):191–217.

G. Alan Marlatt and Dennis M. Donovan. 2005. *Relapse prevention: Maintenance strategies in the treatment of addictive behaviors*. Guilford press.

Vincent Marziano, Tony Ward, Anthony R. Beech, and Philippa Pattison. 2006. Identification of five fundamental implicit theories underlying cognitive distortions in child abusers: A preliminary study. *Psychology, Crime & Law*, 12(1):97–105.

William R. Miller. 2023. The evolution of motivational interviewing. *Behavioural and Cognitive Psychotherapy*, 51(6):616–632.

Amy Phenix, Yolanda Fernandez, Andrew JR Harris, Maaike Helmus, R. Karl Hanson, and David Thornton. 2017. *Static-99R coding rules, revised-2016*. Public Safety Canada.

Miriam Schuler, Hannes Gieseler, Katharina W. Schweder, Maximilian Von Heyden, and Klaus M. Beier. 2021. Characteristics of the Users of Troubled Desire, a Web-Based Self-management App for Individuals With Sexual Interest in Children: Descriptive Analysis of Self-assessment Data. *JMIR Mental Health*, 8(2):e22277.

Michael C. Seto and Angela W. Eke. 2015. Predicting recidivism among adult male child pornography offenders: Development of the Child Pornography Offender Risk Tool (CPORT). *Law and Human Behavior*, 39(4):416–429.

David Thornton, Gwenda M. Willis, and Sharon Kelley. 2024. Dynamic Protective Factors Relevant to Sexual Offending. *Current Psychiatry Reports*, 26(4):142–150.

Fritjof Von Franqué, Ralf Bergner-Koether, Stefanie Schmidt, Jan S. Pellowski, Jan H. Peters, Göran Hajak, and Peer Briken. 2023. Individuals under voluntary treatment with sexual interest in minors: what risk do they pose? *Frontiers in Psychiatry*, 14:1277225.

Tony Ward, Gwenda M. Willis, David S. Prescott, Stijn Vandevelde, Mary Barnao, and Wouter Wanzeele. 2025. *The Good Lives Model of Correctional Rehabilitation: Integrating Theory, Research, and Practice*. Advances in Preventing and Treating Violence and Aggression. Springer Nature Switzerland, Cham.

Gwenda M. Willis and Tony Ward. 2024. Evidence for the Good Lives Model in Supporting Rehabilitation and Desistance from Offending. In Leam A. Craig, Louise Dixon, and Theresa A. Gannon, editors, *The Wiley Handbook of What Works in Correctional Rehabilitation*, 1 edition, pages 299–309. Wiley.

Zixiu Wu, Simone Balloccu, Vivek Kumar, Rim Helaoui, Diego Reforgiato Recupero, and Daniele Riboni. 2023. Creation, Analysis and Evaluation of AnnoMI, a Dataset of Expert-Annotated Counselling Dialogues. *Future Internet*, 15(3):110.

A. Annotation and Agreement

A.1. Consideration of a Scalar Factor Representation

While developing the annotation scheme and guidelines, we explored whether factors could be represented on a scalar continuum rather than as separate risk and protective categories. We decided against this in the present scheme because the relation between risk and protective factors is not fully symmetrical; some protective factors do not have a clear risk-side counterpart, and vice versa. A continuum-based formulation may nevertheless be worth revisiting in future work for selected factor domains.

A.2. Calculation of F_1

The F_1 computation pipeline was implemented in Python 3.10.8. We used dkpro-cassis 0.7.3 for reading INCEpTION XMI files and the corresponding type system, PyYAML 6.0.2 for configuration parsing, and numpy 1.26.4, pandas 2.2.2, and scipy 1.11.4 for metric computation, aggregation, bootstrap-based statistical analysis, and CSV export. F_1 was computed as a greedy span-overlap F_1 with 1:1 matching based on any character overlap of at least one character. We report three variants: *general* (risk vs. protective type only), *exact* (fine-grained factor labels), and *clustered* (fine-grained labels mapped to broader clusters). Scores were aggregated at multiple levels. We computed pairwise F_1 scores within each session and then summarised results as pooled micro-F_1, macro-F_1 across sessions, and macro-F_1 across labels, depending on the analysis. To compare annotation iterations, we estimated differences using session-level bootstrap confidence intervals with 5,000 resamples.

A.3. Exact Factors

Risk factor annotations were dominated by RF7, RF3, RF1, RF4, and RF4.1, while protective factor annotations were most frequent for PF4, PF2, PF8.1, PF4.1, and PF8. Table 6 shows the five most frequent labels per factor type. Table 7 and Table 8 report per-label relaxed *exact* F_1.

A.4. Clustering Factors into Categories

We considered clustering semantically similar fine-grained factors into categories, within their type. For risk factors, we distinguished between *Minor-focused sexual interest and hypersexuality*; *Offence-supportive cognitions and minimisation*; *Affect-/stress-driven coping and impulsivity*; *Intimacy or relationship deficits and social isolation*; *Prior behaviour and learning history*; *Structural instability and young age*; and *Avoidance and/or shame regarding the preference*. Analogously, for protective factors we defined the categories *Sexual responsiveness to adults and non-exclusivity*; *Accurate cognitive appraisal of CSA/CSAM and victim empathy*; *Self-regulation and functional coping*; *Intimacy, social embeddedness and stable relationships*; *Healthy sexual behaviour*; *Structural stability*, understood as satisfaction with housing and employment; and *Acceptance of the sexual preference*. Table 9 shows the clustering of semantically similar fine-grained labels into higher-level categories within their type (RF and PF, respectively). Table 10 reports overall relaxed F_1 for *exact* factors, *general* factor type, and *clustered* categories, indicating

	RF7	RF3	RF1	RF4	RF4.1	PF4	PF2	PF8.1	PF4.1	PF8
$n =$	159	97	85	82	66	163	123	121	119	119

Table 6: Top five most frequent risk and protective factor labels across the pilot corpus.

	PF1	PF2	PF3	PF4	PF4.1	PF5	PF6	PF6.1	PF7	PF8	PF8.1	PF10
i_1	0.72	0.49	0.37	0.58	0.22	0.49	0.07	0.67	0.44	0.25	0.36	0.42
i_2	0.60	0.72	0.29	0.61	0.47	0.44	0.50	0.67	0.80	0.51	0.46	0.00
$\Delta(i_2 - i_1)$	-0.12	0.23	-0.08	0.03	0.25	-0.04	0.43	0.00	0.36	0.26	0.10	-0.42
Support (i_1)	23.5	65.5	24.5	81.0	31.5	22.5	14.5	4.5	20.5	63.0	49.5	12.0
Support (i_2)	15.0	23.5	10.5	33.0	38.0	4.5	2.0	1.5	12.5	19.5	28.0	1.0

Table 7: Per-label relaxed *exact* F_1 for protective factors. Support is the average of label counts across the compared annotations within each iteration.

that categorising factors only slightly improved inter-annotator agreement.

A.5. Confusion and Contrastive Disagreements

We inspected pairwise confusions between fine-grained factor labels. These were concentrated in a small number of conceptually adjacent factors rather than spread broadly across the inventory. The most frequent confusion was PF4.1 vs. RF4 ($n = 5$), followed by PF4 vs. PF4.1, PF4.1 vs. PF8, PF4.1 vs. RF4.1, and RF4 vs. RF7 (each $n = 4$). Overall, the pattern suggests that residual disagreement was primarily contrastive, especially in passages concerning coping and impulse handling. Table 11 shows the most frequent confusions with $n \geq 2$. To further analyse where agreement breaks down, future work could also consider intra-annotator agreement, i.e., the extent to which individual annotators apply the scheme consistently across repeated passes. This would provide additional insight into whether disagreement stems primarily from annotator-specific variation or from ambiguities in the scheme itself.

B. Relapse-Aligned Analyses

B.1. *Protective Ratio* Around Relapse

To supplement the relapse-aligned results reported in Section 6.2, Table 12 provides the corresponding Protective Ratio (PR) values for the pre- and post-relapse windows and their within-client baselines. Sessions were aligned to self-reported relapse, using the session immediately before relapse as the pre-relapse window ($t = -1$) and the session immediately after relapse as the post-relapse window ($t = +1$). Baseline consisted of all other sessions from the same clients with listed relapse events. Differences (Δ) are computed as window minus baseline. For PR, we report both pooled PR and mean session-wise PR. Permutation tests were computed within client.

B.2. Factor-Level Analyses Around Relapse

Tables 13 and 14 list the top five exact-factor movers by absolute effect size for the pre- and post-relapse windows, respectively. For each factor, we report results under two complementary lenses: *normalised rate*, defined as the factor count divided by the total number of annotations in the respective session set, and *presence*, defined as the proportion of sessions in which the factor occurs at least once. All contrasts are computed as window minus baseline. Results are based on within-client permutation tests. The two panels in each table are ranked independently by absolute effect size and are therefore not row-wise matched.

C. Automated Span Extraction

C.1. LLM-based Results

For LLM-based extraction, we experimented with span-based and message-based detection and extraction of protective and risk factors using different LLMs, but achieving the best results with $Qwen2.5{:}14b$[8]. Underlined scores are the ones reported in the main paper.

Table 15 and Table 16 show the detailed results of detecting and classifying protective and risk factors, respectively, without fine-tuning the models. Fine-tuning the model on silver-annotated data for classification improved extracting protective factors, but worsened the extraction of risk factors. Table 17 and Table 18 show the results using the fine-tuned classifier model.

Finally, we also experimented with message-based classification, achieving the best result for

[8] last accessed on 04.10.2025 via https://ollama.com/library/qwen2.5

	RF0	RF1	RF1.1	RF2	RF2.1	RF3	RF4	RF4.1	RF5	RF6	RF7	RF8	RF8.1	RF9	RF10	RF10.1	RF11	RF12
i_1	0.00	0.52	0.00	0.48	0.29	0.47	0.26	0.43	0.57	0.47	0.36	0.29	0.22	0.00	0.62	0.20	0.00	0.50
i_2	0.00	0.69	0.89	0.67	–	0.72	0.36	0.44	0.00	0.40	0.66	0.00	0.40	0.33	0.33	0.17	–	0.00
$\Delta(i_2 - i_1)$	0.00	0.17	0.89	0.19	–	0.25	0.10	0.01	-0.57	-0.07	0.30	-0.29	0.18	0.33	-0.28	-0.04	–	-0.50
Support (i_1)	3.5	53.5	0.5	12.5	24.0	38.5	30.5	34.5	3.5	27.5	56.0	3.5	4.5	2.5	13.0	24.5	1.0	8.0
Support (i_2)	4.5	13.0	4.5	12.0	–	25.0	19.5	9.0	1.5	15.0	48.5	1.5	5.0	6.0	6.0	6.0	–	1.5

Table 8: Per-label relaxed exact F_1 for risk factors. Cells marked – indicate that the label was not present in that iteration. Support is the average of label counts across the compared annotations within each iteration.

	Category	Factors
PF_A	Sexual responsiveness to adults and non-exclusivity	PF1
PF_B	Accurate cognitive appraisal of CSA/CSAM and victim empathy	PF5
PF_C	Self-regulation and functional coping	PF4, PF4.1
PF_D	Intimacy, social embeddedness and stable relationships	PF2, PF7
PF_E	Healthy sexual behaviour	PF10, PF6, PF6.1
PF_F	Structural stability	
PF_G	Acceptance of the sexual preference	PF3
RF_A	Minor-focused sexual interest and hypersexuality	RF1, RF1.1, RF10, RF12
RF_B	Offence-supportive cognitions and minimisation	RF5, RF8.1, RF11
RF_C	Affect-/stress-driven coping and impulsivity	RF10.1, RF4, RF4.1, RF7
RF_D	Intimacy or relationship deficits and social isolation	RF2, RF2.1
RF_E	Prior behaviour and learning history	RF6
RF_F	Structural instability and young age	RF9
RF_G	Avoidance and/or shame regarding the preference	RF3

Table 9: Clustering of fine-grained factors to categories. Some factors remain as singleton categories.

risk factor extraction with this method. Detailed results are shown in Table 19 and Table 20.

The figures in Appendix C.1 show the respective prompts for span and message detection and classification.

C.2. BERT-based Span Extraction (Token Classification)

A list of the BERT models used for experiments can be found in Table 22. The results are presented in Table 21.

	pooled	session-macro micro	session-macro macro
exact	0.45	0.47	0.37
clustered	0.49	0.51	0.40
general	0.59	0.62	0.60

Table 10: Overall relaxed F_1 across all 31 sessions.

Factor A	Factor B	n
PF4.1	RF4	5
PF4	PF4.1	4
PF4.1	PF8	4
PF4.1	RF4.1	4
RF4	RF7	4
RF2	RF7	3
RF0	RF6	3
RF4	RF4.1	2
RF10.1	RF6	2
RF4	RF8.1	2
RF2	RF8.1	2
RF3	RF7	2
PF4	PF8	2
PF2	PF4	2
PF4.1	RF7	2

Table 11: Most frequent pairwise factor confusions. Only confusions with $n \geq 2$ are shown.

Window	Aggregation	PR_{win}	PR_{base}	Δ	p_{perm}
pre ($t = -1$)	pooled	0.439	0.539	-0.101	0.443
pre ($t = -1$)	mean	0.439	0.541	-0.102	0.545
post ($t = +1$)	pooled	0.454	0.521	-0.067	0.640
post ($t = +1$)	mean	0.407	0.565	-0.158	0.317

Table 12: Protective Ratio (PR) in pre- and post-relapse windows compared to within-client baseline. Δ is computed as window minus baseline.

normalised rate			presence		
Factor	Δ	p_{perm}	Factor	Δ	p_{perm}
RF7	0.053	0.157	RF3	0.524	0.266
RF3	0.049	0.321	RF10	-0.413	0.119
PF2	-0.044	0.041	RF7	0.413	0.270
RF1	0.037	0.246	RF1.1	-0.333	0.376
PF4	0.032	0.377	RF8.1	0.317	0.443

Table 13: Top five exact-factor movers by absolute effect size in the **pre-relapse window** ($t = -1$) under the within-client permutation scheme, shown separately for the normalised rate and presence lenses. The two panels are ranked independently by absolute effect size and are therefore not row-wise matched.

normalised rate			presence		
Factor	Δ	p_{perm}	Factor	Δ	p_{perm}
RF4	0.059	0.044	PF8	-0.571	0.058
RF3	0.054	0.286	PF10	-0.413	0.380
PF4	0.047	0.176	PF5	-0.381	0.288
PF4.1	-0.042	0.431	PF6	-0.381	0.466
RF6	-0.041	0.165	RF4	0.333	0.348

Table 14: Top five exact-factor movers by absolute effect size in the **post-relapse window** ($t = +1$) under the within-client permutation scheme, shown separately for the normalised rate and presence lenses. The two panels are ranked independently by absolute effect size and are therefore not row-wise matched.

metric	relaxed match	exact match
precision	0.484	0.481
recall	0.238	0.210
F1	**0.319**	**0.292**

Table 15: Evaluation for span detection and classification on protective factors with $Qwen2.5{:}14b$.

metric	relaxed match	exact match
precision	0.327	0.327
recall	0.149	0.149
F1	**0.205**	**0.205**

Table 16: Evaluation for span detection and classification on risk factors with $Qwen2.5{:}14b$.

metric	relaxed match	exact match
precision	0.431	0.439
recall	0.302	0.290
F1	**0.349**	**0.350**

Table 17: Evaluation for span detection and classification on protective factors with fine-tuned $Qwen2.5{:}14b$.

metric	relaxed match	exact match
precision	0.133	0.132
recall	0.158	0.149
F1	**0.145**	**0.140**

Table 18: Evaluation for span detection and classification on risk factors with fine-tuned $Qwen2.5{:}14b$.

metric	score
precision	0.230
recall	0.383
F1	**0.287**

Table 19: Evaluation for message classification on protective factors with $Qwen2.5{:}14b$.

metric	score
precision	0.272
recall	0.324
F1	**0.296**

Table 20: Evaluation for message classification on risk factors with $Qwen2.5{:}14b$.

Prompt for span detection of protective factors

```
Given the following message, extract text spans that could be potential protec-
tive factors in a therapeutical context.

Instruction:
  • Given the input message, output only the relevant text spans as string,
    nothing else.
  • If multiple Spans can be found, create a list of strings like this:
    ["span1", "span2"]
  • If no span is relevant, output exactly: None
  • Do not add explanations, notes, greetings, or any extra words

Example 1:
input: I can imagine the struggle. Personally, I am finding it rather easy
to engage in a conversation with you
output: I am finding it rather easy to engage in a conversation with you
Example 2:
input: I also joined a choir, go to a theatre improvisation class, see a
friend for coffee every morning (which my wife has a hard time to accept) but
it helps me to gain more confidence in accepting myself
output: ["joined a choir, go to a theatre improvisation class, see a friend
for coffee every morning", "it helps me to gain more confidence in accepting
myself"]
Example 3:
input: Hi!
output: None
Now, classify the following message:
input: {message}
```

Figure 3: Prompt for span detection of protective factors.

Model	Metric	Protective Factors	Risk Factors
		Score	Score
bert-base-cased	Precision	0.002	0.0009
	Recall	0.033	0.010
	Accuracy	0.753	0.693
	F1	**0.004**	**0.0016**
DisorBERT	Precision	0.003	0.0005
	Recall	0.049	0.010
	Accuracy	0.719	0.524
	F1	**0.005**	**0.0009**
MentalBERT	Precision	0.004	0.001
	Recall	0.066	0.010
	Accuracy	0.713	0.738
	F1	**0.007**	**0.0022**

Table 21: Evaluation for token classification on protective and risk factors with fine-tuned BERT models.

Model name	Source	Last Accessed
bert-base-cased	https://huggingface.co/google-bert/bert-base-cased	04.10.2025
DisorBERT	https://huggingface.co/citiusLTL/DisorBERT	04.10.2025
MentalBERT	https://huggingface.co/mental/mental-bert-base-uncased	04.10.2025

Table 22: List of BERT models used for experiments.

Prompt for span detection of risk factors

Given the following message, extract text spans that could be potential risk factors in a therapeutical context.
Instruction:

- Given the input message, output only the relevant text spans as string, nothing else.
- If multiple Spans can be found, create a list of strings like this: ["span1", "span2"]
- If no span is relevant, output exactly: None
- Do not add explanations, notes, greetings, or any extra words

Example 1:
input: Good.. I saw on your questionnaire that you are not in a relationship, and that you are not happy about that. Did I get that right?
output: you are not in a relationship, and that you are not happy about that.
Example 2:
input: Im not good at speaking with other people, due to my anxiety problems 😟
output: ["Im not good at speaking with other people", "my anxiety problems"]
Example 3:
input: Hi!
output: None
Now, classify the following message:
input: {message}

Figure 4: Prompt for span detection of risk factors.

Prompt for span classification of protective factors

You are an annotation system. You must respond **only** with the relevant category labels, exactly as specified. Categories: {protective_factors}
Instruction:

- Given the input span, output only the relevant category labels (e.g., "PF2","PF7"), nothing else.
- If no category applies, output exactly: None
- Do not add explanations, notes, greetings, or any extra words
- One span may fit under multiple categories.

Example 1:
input: Personally, I am finding it rather easy to engage in a conversation with you
output: PF2
Example 2:
input: Just masturbation, sometimes i use legal porn or fantasies of minors (which im trying to cut down on) and i've started using <NAME> chatbots for roleplay - only involving adults
output: PF6,PF10
Example 3:
input: Hi!
output: None
Now, classify the following span:
input: {span}

Figure 5: Prompt for span classification of protective factors.

Prompt for span classification of risk factors

```
You are an annotation system.  You must respond **only** with the relevant cat-
egory labels, exactly as specified.
Categories:
{risk_factors}
Instruction:
  • Given the input span, output only the relevant category labels (e.g.,
    "RF2","RF7"), nothing else.
  • If no category applies, output exactly:  None
  • Do not add explanations, notes, greetings, or any extra words
  • One span may fit under multiple categories.
Example 1:
input:  you are not in a relationship, and that you are not happy about that.
output:  RF2
Example 2:
input:  Im not good at speaking with other people, due to my anxiety problems
output:  RF2,RF7
Example 3:
input:  Hi!
output:  None
Now, classify the following span:
input:  {span}
```

Figure 6: Prompt for span classification of risk factors.

Prompt for message classification of protective factors

```
You are an annotation system.  You must respond **only** with the relevant cat-
egory labels, exactly as specified.
Categories:
{protective_factors}
Instruction:
  • Given the input message, output only the relevant category labels (e.g.,
    "PF2","PF7"), nothing else.
  • If no category applies, output exactly:  None
  • Do not add explanations, notes, greetings, or any extra words
  • One message may fit under multiple categories.
Example 1:
input:  I can imagine the struggle.  Personally, I am finding it rather easy
to engage in a conversation with you
output:  PF2
Example 2:
input:  Just masturbation, sometimes i use legal porn or fantasies of minors
(which im trying to cut down on) and i've started using <NAME> chatbots for
roleplay - only involving adults
output:  PF6,PF10
Example 3:
input:  Hi!
output:  None
Now, classify the following message:
input:  {message}
```

Figure 7: Prompt for message classification of protective factors.

Prompt for message classification of risk factors

```
You are an annotation system.  You must respond **only** with the relevant cat-
egory labels, exactly as specified.
Categories:
{risk_factors}
Instruction:
  • Given the input message, output only the relevant category labels (e.g.,
    "RF2","RF7"), nothing else.
  • If no category applies, output exactly:  None
  • Do not add explanations, notes, greetings, or any extra words
  • One message may fit under multiple categories.
Example 1:
input:  Good..  I saw on your questionnaire that you are not in a relationship,
and that you are not happy about that.  Did I get that right?
output:  RF2
Example 2:
input:  Im not good at speaking with other people, due to my anxiety problems
😟
output:  RF2,RF7
Example 3:
input:  Hi!
output:  None
Now, classify the following message:
input:  {message}
```

Figure 8: Prompt for message classification of risk factors.