A Retrieval-Augmented LLM Pipeline for Generating Cognitive Conceptualization Knowledge Graphs

Simone Pinna[1,†], Silvia Maria Massa[1,†], Diego Reforgiato Recupero[1,†] and Daniele Riboni[1,*,†]

[1]Department of Mathematics and Computer Science, University of Cagliari, via Ospedale 72, 09124 Cagliari

Abstract

Cognitive Conceptualization Diagrams (CCDs) are commonly used in Cognitive Behavioral Therapy (CBT) for representing the relationships among a patient's personal history, beliefs, thoughts, emotions, and behaviors. Currently, CCDs are manually authored by experts, which limits the availability of structured resources for research and training. In this paper, we introduce an LLM-based framework for (i) automatically generating inferred CCDs from brief patient descriptions and psychotherapy transcripts, and (ii) representing them as Knowledge Graph (KG) that can be inspected and queried. Our method adopts a Retrieval-Augmented Generation (RAG) module, which selects semantically related transcripts from a real-world corpus to produce a CBT-oriented psychological report. We use prompt-guided LLM extraction and hierarchical machine learning classification of core beliefs to produce a detailed CCD from the psychological report. We further formalize CCDs as an RDF-based semantic model enabling structured semantic representation. Using expert-created CCDs as gold references, we generated a dataset of 99 CCDs for different patient typologies, and evaluated their quality considering semantic similarity, contradiction analysis, and agreement on major and fine-grained core-belief labels. We also performed a qualitative assessment using an LLM-as-a-judge procedure grounded in established CBT formulation criteria. Results show that our method produces CCDs that are semantically and structurally consistent with expert formulations, improves fine-grained core belief classification over different baselines, and achieves the strongest qualitative ratings among compared pipelines. These findings suggest that our system can support scalable and clinically meaningful structuring of psychotherapy case formulations. The cognitive conceptualization KG is publicly available online.

Keywords

Cognitive Behavioral Therapy, Cognitive Conceptualization Diagrams, Large Language Models, Retrieval Augmented Generation, Knowledge Graphs

1. Introduction

Over the last decade, there has been an increase in mental health disorders, especially among younger people [1]. Mental health problems are becoming common and can have serious consequences, such as suicide, which has become the second leading cause of death among young people aged 10-24 [2]. Between 2011 and 2017-18, rates of adolescent depression in the U.S. increased by at least 60% [3]. This problem has increased the interest toward mental well-being, with the percentage of mental healthcare utilization in U.S. growing from 20% in 2019 to 23% in 2022 [4]. However, access to mental health services faces several obstacles, including the limited number of therapists, geographical barriers, and costs [5]. To address these issues, the field of Cyber Health Psychology has emerged, which aims to enhance healthcare delivery and patient engagement in the digital age [6]. Intervention through digital platforms has been demonstrated to provide benefits by smoothing out some of these barriers [7].

In particular, several studies indicate that Artificial Intelligence (AI) may be a valuable tool in the field of mental health care [6, 8, 9]. Recently, Large Language Models (LLMs) showed promising results in predicting mental health disorders [10] and supporting diagnosis [11]. Improvements have been

ESWC'26: 5th Workshop on LLM-Integrated Knowledge Graph Generation from Text (TEXT2KG), May 10-14, 2026 | Dubrovnik, Croatia.

*Corresponding author.

†These authors contributed equally.

simone.pinna2@unica.it (S. Pinna); silviam.massa@unica.it (S. M. Massa); diego.reforgiato@unica.it (D. R. Recupero); riboni@unica.it (D. Riboni)

0009-0007-7141-2407 (S. Pinna); 0000-0002-3285-8971 (S. M. Massa); 0000-0001-8646-6183 (D. R. Recupero); 0000-0002-0695-2040 (D. Riboni)

suggested, such as the implementation of Retrieval Augmented Generation (RAG) techniques [12], fine-tuning on a variety of datasets [10], and the integration of cognitive models into an LLM prompt [13].

Different works in this field target specific approaches, such as Cognitive Behavior Therapy (CBT) [14]. CBT is an umbrella term for therapeutic techniques based on the assumption that psychopathology is the product of faulty information processing that manifests itself in distorted and dysfunctional thinking, which directly leads to negative emotions and maladaptive behaviors. In the CBT approach, Cognitive Conceptualization Diagrams (CCDs) represent a structured and defined way to represent patients cognitively [15]. Recently, the use of CCDs within LLM prompts has been proposed to support psychotherapists' training [13]. In that work, experts manually created a set of fictional CCDs based on hypothetical patient profiles. However, because this process was manual and required domain expertise, scaling it to address a broader range of psychological profiles would be expensive. This limitation is common in mental health research, where the use of public knowledge bases is particularly challenging because of the sensitive and private nature of the underlying information. As a result, publicly available datasets are often small and biased, which limits model performance and generalization [16].

Considering the value of CCDs for different mental health applications, and the lack of publicly available structured datasets for CBT [17], in this paper we propose an LLM-based framework for the automatic creation of inferred CCDs from brief patient descriptions expressed in natural language. The proposed framework follows a RAG-like structure that uses a real-world public dataset of therapy session transcripts as its knowledge base. The pipeline relies on an LLM with a specific prompt and a two-step refinement process, including fine-grained belief classification, to create a detailed CCD corresponding to the patient description. Moreover, we propose a semantic knowledge model for CCDs, formalized using the Resource Description Framework (RDF). The model describes the relationships between CCD fields, reflecting the chain of causality defined in [18]. The scheme applied to the generated CCDs allows them to be represented as a KG that can be inspected and queried to support different applications in the CBT domain.

We performed an extensive experimental evaluation of our framework, using the fictional CCDs defined by experts in [13] to derive patients descriptions. The effectiveness of our technique is evaluated considering similarity metrics between the CCDs defined by experts and those automatically derived using our framework. We also performed a qualitative evaluation based on the scale of values proposed in [19] using an LLM-as-a-judge approach. The obtained results show that our proposed framework produces CCDs that are semantically and structurally similar to the reference models. The qualitative evaluation favors the CCDs produced by our framework with respect to different baselines, showing that the use of therapy session transcripts improves the CCD accuracy.

In summary, the main contributions of this work are the following:

- We introduce a novel framework for deriving detailed CCDs starting from a brief patient's description. Our pipeline uses a RAG approach, exploits a dataset of psychotherapy session transcripts, and includes fine-grained belief classification and two generative steps.
- We formalize an ontology for the derived CCDs in the form of a KG, in order to make explicit the relationships between the CCD fields, reflecting the principles of CBT.
- We created and publicly released a dataset of 99 CCDs inferred by our framework based on the patient's descriptions and histories defined by experts in [13]. The dataset is publicly available through the Virtuoso platform at: https://192.167.149.12:9001/sparql/.
- We performed a detailed experimental evaluation including both quantitative and qualitative metrics, which shows the potential of our approach.

The rest of the paper is structured as follows. Section 2 discusses related work. The LLM-based pipeline for CCD derivation is presented in Section 3. The experiments conducted for evaluation are described, and the results are reported and discussed in Section 4. Finally, Section 5 summarizes the work carried out, presents the current limitations and future developments.

2. Related Work

In this section, we provide background information regarding CBT and CCDs, their adoption in digital technologies, and existing works about knowledge representation in psychotherapy applications.

2.1. Cognitive Behavioral Therapy and Cognitive Conceptualization Diagram

CBT was introduced by Dr. Aaron Beck between 1960 and 1970, and is based on the theory that a person's thoughts and beliefs influence their emotions and behaviors [18]. It was the first psychotherapy approach largely identified as evidence-based in most clinical guidelines, and is now one of the most widely adopted frameworks in the field of psychotherapy [20]. The CCD is a tool used to organize the multitude of patients' characteristics, providing a cognitive map of the patient's psychological profile. The diagram is defined based on CBT therapy sessions with the patient, and is continuously refined as the therapist learns new information. A diagram is composed of information such as Automatic Thoughts (ATs), emotions, behavior, and/or beliefs, which interact and influence each other. Judith S. Beck defined a framework consisting of the following fields: Relevant Childhood Data, Core Belief(s), Conditional Assumptions/Beliefs/Rules, Compensatory/Coping Strategies, Situation, AT, Meaning of the AT, Emotion, Behavior. Of particular importance are Core Beliefs, i.e., ideas about themselves, other people, and their world that are so deeply rooted that they are often unconscious. A. Beck and J. S. Beck provide a standard list of negative Core Beliefs, divided into three major categories: helpless, unlovable, and worthless [18]. To address the lack of methods for evaluating CCDs, a qualitative value scale has been proposed that compares cognitive models with reliability and quality criteria, focusing on what defines a "good enough" formulation [19]. The scale is a useful tool for evaluating self-generated cognitive models and has therefore also been used for the qualitative evaluation of the framework proposed in this work. Other case formulation frameworks include the one proposed by J. B. Persons, which is centered around four components: problems and symptoms, psychological mechanisms, precipitants, and origins [21]. Unlike the CCD, this model does not explicitly represent the causal chain between beliefs, emotions, and behaviors.

2.2. Digital Technologies in CBT

Due to the relevance of CBT in psychotherapy, various studies aim to implement its principles in Cyber Health Psychology to achieve different goals. The most common use of CBT techniques in technology is the creation of chatbots for psychotherapy; for example, out of 17 chatbots offering psychotherapy, 10 were based on CBT [22]. Integration of AI into CBT practice is implemented in various stages: pre-treatment, therapeutic process, and post-treatment [17]. The creation of datasets in the field of CBT mainly concerns the identification of Cognitive Distortions [23], Core Beliefs [24] and understanding knowledge [24] [25]. AI is also used for predictive purposes on individual aspects of a CBT framework. For instance, Furukawa et al. proposed the use of Google T5 to link a specific emotion to an AT [26]. In our work, we use different machine learning models to derive major and fine-grained Core Beliefs.

The use of CCDs provides a formal data structure for use within an LLM prompt. In [13], this method is used to simulate patients with the aim of training psychotherapists to correctly recognize the key principles of CBT. The diagrams are created by experts in the field based on transcripts of psychotherapy sessions in the Alexander Street database. To the best of our knowledge, it is the only existing dataset containing complete CCDs derived from real patients psychotherapy sessions and created by clinical psychologists. The resource is not freely accessible, but can be requested for research purposes, and provides 106 cognitive models, often relating to the same patient subjected to different situations. The goal of our system is to generate scalable, evidence-grounded CCDs from patients descriptions while preserving the structural and clinical principles of CBT, without requiring expert manual effort.

2.3. Knowledge representation in psychotherapy

The structured representation of information has been shown to improve the identification of meaningful relationships and also shows great potential for integration with AI [27]. KGs are widely used in medicine [28]: large-scale efforts such as SNOMED-CT [29], the Human Disease Ontology [30], and the NCI Thesaurus [31] provide standardized, interoperable vocabularies for representing clinical concepts and their relationships. More recent work has explored the automatic construction of KGs from electronic health records [32], showing the potential of data-driven approaches for clinical knowledge representation. In psychology, the available resources are still limited, despite showing great potential [33]. Some KGs have been applied to mental health research, especially for population-level or disease-centric representations, such as disorder ontologies and biomedical relation graphs [34, 35]. The CCD KG proposed in this work differs from those approaches since it represents patient-specific cognitive formulations, rather than population-level clinical knowledge or diagnostic categories.

KGs can provide a semantic framework of relationships for complex mental health data, and are used together with LLMs to overcome their limitations [36]. In CBT, the use of formal data structures is limited. The only existing structured resource is an ontology modeling the theoretical axioms of CBT for depression [37], which operates at the level of clinical theory rather than individual patient cases. The structured representation of CBT principles can be used to show non-trivial relationships between thinking errors, emotions, and situations, and to predict cognitive distortion [38]. However, no structured representation exists for CCDs. This leaves a gap between the unstructured natural language used during psychotherapy sessions and a structured semantic formalization of key CBT concepts at the individual patient level.

3. Methodology

This section describes the proposed pipeline for deriving a CCD from the patient's textual description. The inferred CCDs are represented using a semantic model, constituting a KG. The pipeline is shown in Figure 1. It is divided into two main modules: the RAG module, and the CCD module, each consisting of different phases. The **User Input** is the description of the patient, consisting of their history and current situation. **The Data Source Input** is a subset of the Alexander Street digital archive present on Kaggle. The dataset used consists of 2,022 transcripts of real psychotherapy sessions between a patient and a therapist, enriched with metadata. The **Output** of the pipeline is the CCD represented in RDF format, including the main conceptual components of the model: History, Core Beliefs, Intermediate Belief, Intermediate Belief Depression, Coping Strategies, Situation, Auto Thought, Emotion, Behavior. The inferred CCDs constitute a KG that can be queried and inspected for different mental health applications.

3.1. RAG Module

The RAG module is defined as such because it combines a phase of retrieval of information semantically related to the user's query with a phase of LLM generation based on that information. The purpose of the module is to return a complete psychological report based on the transcript of a patient's psychotherapy session that is consistent with the description provided by the user. In the **Retrieval Similarity Comparison** phase, the transcript most similar to the patient's description is chosen using a comparison method based on the cosine similarity between the embeddings of the description and the embeddings of the summaries of each transcript.

The knowledge base used is the **psychotherapy session transcripts dataset**[1], summarized using an LLM. Typically, full transcripts are several thousand tokens long, while the patient descriptions are a few hundred tokens long. Since comparing embeddings of texts that are deeply unbalanced in size may bias similarity estimates and reduce retrieval accuracy, the asynchronous **LLM transcript summary** phase is introduced to summarize all of the transcripts in the dataset. Considering the size of the full transcripts, summaries are produced by dividing the text into fixed-size chunks and then

[1] https://www.kaggle.com/datasets/abdurrasoolphd/psychotherapy-transcripts

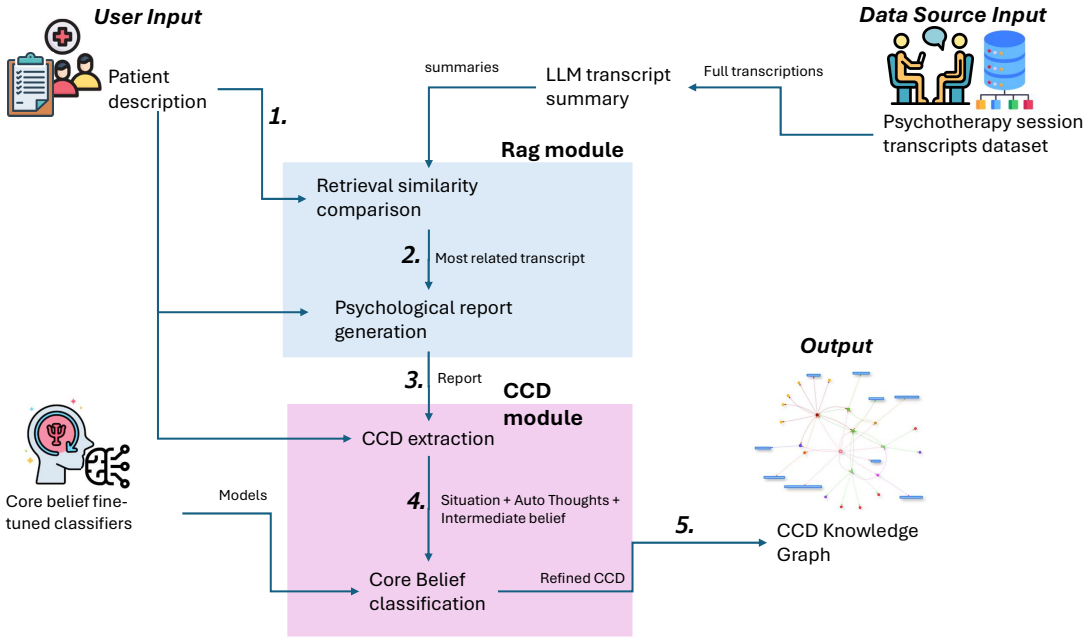

Figure 1: Workflow of the proposed pipeline for automated CCD generation. (1) The patient description compared against a dataset of psychotherapy transcripts summarized via retrieval similarity; (2) the most related transcript is selected; (3) a psychological report is generated and passed to the CCD module; (4) CCD components, Situation, Auto Thoughts, and Intermediate Beliefs, are extracted and refined through fine-tuned Core Belief classifiers; (5) the inferred CCD represented in RDF format is produced as output.

concatenating them. Retrieval also uses a fallback method in case no transcript is sufficiently similar; in that case, a similarity check is performed with a series of metadata associated with the dataset element. The metadata used in that case are: Real Title (i.e., the title of the transcription in the dataset), Psyc Subjects (the manually classified subjects), and Symptoms annotated in the dataset.

The most relevant transcript is given as input to the **Psychological Report Generation** phase. The aim of this phase is to use an LLM to analyze the transcript and generate a document defined as a Psychological Report (PR) that outlines the psychological profile that emerged during the session. The PR is not intended to be a summary, but rather a CBT-oriented analytical document. The prompt used for this end defines the aspects most relevant to CBT, such as thoughts and behaviors, asking for an explicit motivation for them based on the text. The PR is a knowledge base rooted in real psychological facts and motivated by a psychotherapy session. The following prompt was used to generate the PR. Minor phrasing adjustments were made for readability.

> *You are an expert clinical psychologist with strong experience in Cognitive Behavioral Therapy (CBT) Beck style.*
>
> *TASK: Write a psychological report based on the DESCRIPTION of the patient and TRANSCRIPTION of SIMILAR (but DIFFERENT) THERAPY session, with the following constraints.*
>
> *STRICT FACTUAL CONSTRAINTS:*
> *- The ONLY factual events, situations, and life circumstances you are allowed to mention are those explicitly listed in the DESCRIPTION.*
> *- You MUST NOT introduce any new facts, events, diagnoses, medications, incidents, relationships, or life details that are not present in the DESCRIPTION.*
> *- The TRANSCRIPT DOES NOT describe the same patient and MUST NOT be used as a source of factual information.*
>
> *USE OF THE TRANSCRIPT:*
> *- The transcript may ONLY be used as a source of psychological patterns and mechanisms (e.g., typical emotional reactions, thought styles, behavioral responses, coping tendencies).*
> *- You may use the transcript ONLY to support or justify psychological interpretations of the DESCRIPTION*

facts.
- Never state or imply that an event described in the transcript happened to the patient.

PSYCHOLOGICAL INTERPRETATION RULES:
- Every psychological interpretation must be explicitly grounded in one or more facts from the DESCRIPTION.
- You may infer emotional states, cognitive tendencies, behavioral responses, and coping styles, but ONLY as psychological reactions to the described facts.
- Do NOT speculate about unmentioned traumas, incidents, medical conditions, legal issues, or future plans.

STYLE REQUIREMENTS:
- Write in a CBT case-formulation like Beck-style therapy notes: concrete, descriptive, focused on the patient's inner experience (thoughts, emotions, behaviors), with minimal theoretical language.
- Focus on how a person with these DESCRIPTION facts might psychologically experience them.
- Avoid introducing concrete details not present in the DESCRIPTION.

CONTENT GOAL: Produce a detailed psychological report that:
- Explains the likely emotional, cognitive, and behavioral impacts of the DESCRIPTION facts.
- When describing behaviors or coping, use concrete action verbs (e.g., withdraws, avoids, ruminates, procrastinates...).
- When describing emotions, prefer clustered, everyday labels similar to: "anxious, worried, fearful"; "sad, down, lonely".
- Uses the TRANSCRIPT only as indirect psychological support.
- Remains strictly faithful to the DESCRIPTION as the source of facts.

DESCRIPTION: clean-description

TRANSCRIPT: clean-transcript

3.2. CCD Module

The CCD Module uses the patient's description, PR, and Core Belief classifiers to generate CCDs that are consistent with the input and reflect the techniques used in CBT. The first phase of **CCD extraction** aims to derive the CCD fields using the concrete facts present in the description, adding psychological depth by referring to the PR. Specifically, the situation and history fields are transcribed as faithfully as possible, while the others are generated around them, without contradicting them, but trying to motivate the patient's behaviors and thoughts. Generation is performed using LLM with a carefully designed prompt based on knowledge from "Cognitive behavior therapy: Basics and beyond" [18], the reference book for modern CBT. This work studies the ability of an LLM to create a CCD when guided with one of the most comprehensive and relevant references on the subject. The prompt is divided into three sections: role and rules to be followed, JSON structure of the output, and definitions of each individual field. The fields and their meanings are shown in Table 1; the relationships between them will be discussed in Section 3.3. The rules also specify a subsection called BECK-STYLE SELF-QUESTIONS, which lists common questions that a CBT therapist asks to identify the typical patterns of the fields. The prompt is structured as reported below with minor paraphrasing. This is the prompt referred to in Section 4.2 as 'Our Prompt', on which the other methods are based.

You are a senior CBT clinician and case formulator (Beck-style).

You are given: 1) A DESCRIPTION of concrete facts about the patient case. 2) A complete REPORT with facts and psychological evidence about a transcription of a SIMILAR patient case but not the SAME.

TASK: Construct a CBT Cognitive Conceptualization Model (CCD-like) for the CLIENT, asking yourself key case-formulation questions using the REPORT. The facts have to remain the same of the DESCRIPTION.

ABSOLUTE RULES:
- ADD to every field the psychological aspects but DON'T CHANGE the real facts especially history and situation.
- Psychological evidence means: the client's reported thoughts, emotions, behaviors or patterns.
- You MUST keep the JSON schema provided.

BECK-STYLE SELF-QUESTIONS (use transcript evidence to answer internally before producing JSON):
- What is the main real-life triggering situation the client is dealing with?
- When the client deals with the situation, what was going through their mind (hot thought)?
- What did that thought mean about them / others / the future?
- What emotions followed, and what did they do next (behavior)?
- What rule/assumption would generate that thought in that situation (intermediate belief)?
- What coping strategies keep the cycle going?

DEFINITIONS: <definitions of each field>

OUTPUT: Return ONLY valid JSON, with EXACTLY the following keys and structure (no extra keys).

JSON SCHEMA: <fields in JSON formats>

DATA YOU RECEIVE:
1) DESCRIPTION: {description}
2) REPORT: {report}

Field	Meaning
History	Relevant background the client reports that helps explain the development and/or maintenance of difficulties.
Core Beliefs (CB)	Deeply held, global, rigid, overgeneralized beliefs about self, others, and the world.
Intermediate Belief (IB)	Rules, assumptions, and attitudes that link beliefs to how the person interprets and navigates real-world situations.
Intermediate Belief Depression (IBD)	A depression/hopelessness-leaning variant of the intermediate belief.
Coping Strategies (CS)	Recurring strategies the client uses to manage distress, especially those that unintentionally maintain problems.
Situation	Real-life triggering context.
Automatic Thought (AT)	A thought that pops up immediately in the moment as a response to the situation.
Emotion	The emotional reaction that follows the automatic thought(s).
Behavior	The concrete behavioral response to the situation.

Table 1
The generated CCD fields and their corresponding meanings in CBT.

The generated CCD is further refined using two Transformer-based classifiers, dedicated respectively to the classification of Core Beliefs Major (CBsM) and Core Beliefs Fine-Grained (CBsFG) fields. The models, initially pre-trained on generic corpora, were subsequently fine-tuned for multi-label classification tasks. The datasets used for fine-tuning are CBT-PC[2] and CBT-FC[3] , which are part of the broader CBT-Bench benchmark [24]. The two datasets contain 184 and 112 examples, respectively, consisting of the original text, the situation, the thought that arises in response, and the CBsM or CBsFG label assigned by expert annotators. Table 2 shows the 3 CBsM labels available with their 19 associated CBsFG labels, using the naming conventions defined by Beck [18]. Since multiple CBsM and CBsFG can coexist in the same individual, the problem is modeled as two independent multi-label classification tasks, taking into account the hierarchy during inference. Predictions are obtained by applying a threshold to the probabilities produced by the classifier, rather than by selecting the class with the highest probability. The thresholds were chosen empirically using F1 micro as the metric to maximize on a validation set (80/20 split of the training data). For the Major labels, a threshold of 0.5 was applied to helpless and unlovable, while 0.6 was used for worthless, which showed greater sensitivity to false

[2]https://huggingface.co/datasets/Psychotherapy-LLM/CBT-Bench/viewer/core_major_test
[3]https://huggingface.co/datasets/Psychotherapy-LLM/CBT-Bench/viewer/core_fine_test

positives at lower thresholds; for the fine-grained labels, a threshold of 0.30 was applied, yielding the best performance across classes. The two classifiers are used in the **Core Belief Classification** phase, which takes the Situation, AT and IB fields as input. Initially, the CBsM classifier is used, followed by the CBsFG classifier, which can only return labels belonging to the relevant active CBsM, thus ensuring the hierarchical consistency of the predictions. The CBsM and CBsFG hypothesized by the LLM are replaced with the predicted labels, and the refined CCD is returned as the module's output.

Core Beliefs Major	Core Beliefs Fine-Grained
helpless	I am incompetent; I am helpless; I am powerless, weak, vulnerable; I am a victim; I am needy; I am trapped; I am out of control; I am a failure, loser; I am defective.
unlovable	I am unlovable; I am unattractive; I am undesirable, unwanted; I am bound to be rejected; I am bound to be abandoned; I am bound to be alone.
worthless	I am worthless, waste; I am immoral; I am bad, dangerous, toxic, evil; I don't deserve to live.

Table 2
Core Belief Major categories and their associated Core Belief Fine-Grained items

3.3. CCD Knowledge Graph

The CCDs generated by the proposed pipeline are represented in JSON format, making them available in a semi-structured form. An example is shown in Figure 2(a). However, JSON representation does not include the causal chains between the CBT fields. Hence, we introduce a CCD ontology that describes their semantic structure. The ontological schema applied to individual data instances generates a KG composed by 99 CCDs represented as RDF graphs serialized in Turtle (ttl) format and stored in a triple store. The representation of the CCD is shown in Figure 2(b), which illustrates an extract of some relationships. The main relationships are the following. The patient's history forms the personal Core Beliefs, which influences the development of an IB, manifested as AT in response to a given situation. The AT in reaction to the triggering situation generates a response from both a behavioral (behavior) and emotional (emotion) point of view. The emergence of IBs and IBDs leads to the development of the patient's defense through the use of Coping Strategies.

The modeled ontology comprises eleven classes, twenty object properties, and two data types. The CCD in JSON format, refined by the Core Belief classification, is represented by applying the ontological schema described through the instance of the entities provided by the model (e.g., CaseFormulation, Situation, AutomaticThought, IntermediateBelief, ActiveCoreBelief). These entities are linked through object properties that model the conceptual relationships between the elements of the CCD (e.g., hasSituation, givesRiseTo, underlies), while textual information and identifiers are represented through datatype properties. In this way, the narrative structure of the CCD is transformed into an explicit formal representation. The ability to query CCDs and their individual fields allows the identification of non-trivial patterns and the discovery of new, seemingly invisible relationships. The complete ontology schema is available both graphically and as a Turtle file at: https://anonymous.4open.science/r/ccd-ontology-schema-5DE7.

Figure 3 shows an example of a SPARQL query that identifies patients with a history of substance abuse who have developed a helpless Core Belief.

4. Experimental evaluation

This section describes the experiments conducted, the used metrics, and the achieved results.

```
{
 "id": "1-1",
 "history": "The client has a history of substance abuse and has
completed rehabilitation, indicating a significant emotional journey.
He reports ongoing family difficulties, especially with his mother, ...",
 "helpless_belief": ["I am helpless", "I am defective", "I am trapped",
"I am out of control", "I am powerless, weak, vulnerable"],
 "unlovable_belief": ["I am unlovable", "I am bound to be abandoned" ],
 "worthless_belief": [],
 "intermediate_belief": "If I put myself in social situations,
people will judge or reject me because of my weight and past.
It's safer to avoid these situations than risk feeling humiliated or
unwanted...
 "intermediate_belief_depression": "No matter what I do, nothing will
change..."
 "coping_strategies": "He avoids responding to invitations and withdraws
from family
 contact to manage anxiety and protect himself from anticipated rejection
or negative evaluation ...
 "situation": "Received an invitation to a cousin's upcoming wedding. Has
not responded
 to the RSVP request and has not answered follow-up calls from family
members.",
 "auto_thought": "If I go, everyone will judge me for my weight and think
I'm a failure.
 It's better if I just don't show up so I don't have to face their rejection.",
 "emotions": ["anxious", "ashamed", "embarrassed", "sad", "lonely",
"fearful" ],
 "behavior": "He ignores the RSVP request and avoids answering follow-up
calls from family members. He withdraws from communication and does
not engage with the invitation."
}
```


(a) CCD example representing one case formulation.


```
...
ex:case_1_1 a ex:CaseFormulation ;
  ex:externalId "1-1" ;
  ex:hasActiveCoreBelief ex:activeCoreBelief_1_1_helpless,
    ex:activeCoreBelief_1_1_unlovable ;
  ex:hasAutoThought ex:autoThought_1_1 ;
  ex:hasBehavior ex:behavior_1_1 ;
  ex:hasCopingStrategies ex:coping_1_1 ;
  ex:hasHistory ex:history_1_1 ;
  ex:hasIntermediateBelief ex:intermediateBelief_1_1 ;
  ex:hasIntermediateBeliefDepression ex:intermediateBeliefDep_1_1 ;
  ex:hasSituation ex:situation_1_1 ;
ex:CB_helpless a ex:CoreBeliefCategory .
ex:CB_unlovable a ex:CoreBeliefCategory .
ex:EM_anxious a ex:Emotion ;
  rdfs:label "anxious" .
...
ex:behavior_1_1 a ex:Behavior ;
  ex:behavioralReactionTo ex:autoThought_1_1 ;
  ex:text "He ignores the RSVP request and avoids answering follow-up calls
from family members. He withdraws from communication and does not
engage with the invitation." .
ex:coping_1_1 a ex:CopingStrategySet ;
  rdfs:label "Coping strategies" ;
  ex:copingResponseToBelief ex:intermediateBeliefDep_1_1,
    ex:intermediateBelief_1_1 ;
  ex:text « " .
ex:intermediateBelief_1_1 a ex:IntermediateBelief ;
  ex:givesRiseTo ex:autoThought_1_1 ;
  ex:isInfluencedBy ex:activeCoreBelief_1_1_helpless,
    ex:activeCoreBelief_1_1_unlovable ;
  ex:text "If I put myself in social situations, people will judge or reject me
because of my weight and past..." .
...
```


(b) Extract of RDF representation of the CCD example.

Figure 2: Example of cognitive case conceptualization represented as (a) a CCD diagram in JSON and (b) its RDF KG encoding. Ellipses (...) indicate that additional elements and relationships are present but omitted for brevity.

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX schema: <http://schema.org/>
SELECT DISTINCT ?patientId
WHERE {

    ?case a ex:CaseFormulation ;
          ex:externalId ?patientId ;
          ex:hasHistory ?history ;
          ex:hasActiveCoreBelief ?coreBelief .
    ?history ex:text ?historyText .
    FILTER(CONTAINS(LCASE(?historyText), "substance abuse"))
    ?coreBelief a ex:ActiveCoreBelief ;
                ex:coreCategory ex:CB_helpless .
}
```

Figure 3: Example query that retrieves the IDs of patients who have “substance abuse” in their history and the “helpless” Core Belief active.

4.1. Metrics

The effectiveness of a generated CCD can be verified by its similarity to a CCD defined as ‘gold’ produced by experts in the field [39]. Hence, our evaluation was conducted on three different levels: similarity with the CCDs produced by experts, as well as the CCD quality, and the quality of the Core Beliefs classification. In the following, we report the metrics used for each of these levels.

Semantic Textual Similarity (STS) assesses the degree to which two sentences are semantically equivalent to each other [40]. This task serves to evaluate how semantically similar the fields of the generated CCDs are to the fields of the CCDs produced by experts, based on the patient’s history and situation. The four metrics used to assess STS are:

- Embeddings Cosine Similarity: Cosine Similarity quantifies semantic similarity between dense

vector embeddings, becoming a very popular measure in STS tasks [41]. It is calculated as the normalized cosine of the angle between two vectors. Values closer to 1 indicate greater semantic similarity.

- CrossEncoder STS: a model that evaluates semantic similarity by processing two sentences together as a single input. In this way, the model can directly compare the tokens of the two sentences and produce a normalized similarity score in [0,1], where 1 indicates the maximum similarity [42].
- METEOR: a lexical similarity metric that evaluates the degree of overlap between a hypothesized text and a reference text through word-by-word alignment. Unlike metrics based purely on n-grams, METEOR integrates exact matches, morphological variants and synonyms, combining precision and recall into a single score [43]. METEOR produces a score in the range [0, 1], where higher values indicate greater lexical similarity between the hypothesized and reference text.
- NLI Contradiction: Natural Language Inference (NLI) is the task of identifying the relationship between a pair of sentences, premise and hypothesis, as implication, neutrality or contradiction. Implication occurs if the premise implies the hypothesis; contradiction occurs if the hypothesis contradicts the premise; neutrality occurs in the remaining cases [44]. NLI models return the probability of the three labels. In this work, the average between the probability that the gold CCD field (premise) contradicts the predicted CCD field (hypothesis) and vice versa is used. The probability of implication was not considered because the addition of psychologically relevant details invalidates the strict logical implication despite preserving semantic consistency. The goal of this metric is therefore to evaluate whether the LLM generates CCD fields that remain semantically consistent with the reference standard while allowing the introduction of relevant complementary information. The value is in the range [0, 1], with lower values indicating a lower probability of contradiction.

These quantitative metrics offer an objective way to assess whether the generated models are semantically similar to the models created by experts. However, a CCD could be particularly similar to the reference CCD without being a high-quality CCD. We adopt this definition of quality proposed in [19]: '*a parsimonious, coherent and meaningful account of a client's presenting problems in cognitive therapy terms. This refers to whether the formulation is coherently structured, with the elements within the formulation interacting with the presenting problems in a meaningful manner across situations or time*'. For this reason, we adopt the Quality of Cognitive Case Formulation Rating Scale, defined as a way of establishing the quality of cognitive therapy case formulations. The scale is built around the definition of what makes a case formulation 'good enough'; i.e., a formulation that includes a majority of relevant information (childhood data, core beliefs, dysfunctional assumptions and compensatory strategies) at an appropriate level of detail and links this information to prototypical problematic situations. On the contrary, a 'Very Poor' formulation shows minimal integration of the elements, includes irrelevant information, and shows evidence of misunderstanding the CCD. The complete Ordinal Rating Scale is: 'Good', 'Good Enough', 'Poor', 'Very Poor'.

The agreement between quantitative and qualitative assessments was evaluated by calculating Cliff's delta (Δ), which quantifies the degree of stochastic dominance between two distributions. Cliff's delta represents the difference between the probability that an observation from one group is greater than an observation from the other group and the inverse probability, considering all possible comparisons between the two distributions [45]. It is useful when data are non-normal, or are ordinal and hence have reduced variance, as in the case of this work, where an ordinal scale is used for quality and similarity metrics have no guarantee of normal distribution. The coefficient varies in the range $[-1, +1]$, where 0 indicates complete overlap between the distributions, while values close to the extremes indicate no overlap and dominance of one group over the other.
It is defined as:

$$\Delta = P(X > Y) - P(X < Y) \tag{1}$$

And calculated as:

$$\Delta = \frac{\#(x_i > y_j) - \#(x_i < y_j)}{n_X n_Y} \tag{2}$$

where $X = \{x_1, \ldots, x_{n_X}\}$ and $Y = \{y_1, \ldots, y_{n_Y}\}$ are the two samples, n_X and n_Y are their sample sizes, and $\#(x_i > y_j)$ ($\#(x_i < y_j)$) denotes the number of pairwise comparisons (x_i, y_j) such that x_i is greater (smaller) than y_j, across all $n_X n_Y$ pairs. Ties ($x_i = y_j$) do not contribute to the numerator.

Both Major and Fine-Grained Core Beliefs classifiers were evaluated using standard multi-label classification metrics, chosen to capture different aspects of model performance:

- F1 micro (F1 μ): aggregates the contributions of all classes before computing the harmonic mean of precision and recall. It is dominated by frequent classes and reflects overall token-level correctness across the entire label set.

$$F1_{micro} = \frac{2 \cdot \text{Precision}_{micro} \cdot \text{Recall}_{micro}}{\text{Precision}_{micro} + \text{Recall}_{micro}} \tag{3}$$

  where:

$$\text{Precision}_{micro} = \frac{\sum_{c \in \mathscr{L}} TP_c}{\sum_{c \in \mathscr{L}} (TP_c + FP_c)} \tag{4}$$

$$\text{Recall}_{micro} = \frac{\sum_{c \in \mathscr{L}} TP_c}{\sum_{c \in \mathscr{L}} (TP_c + FN_c)} \tag{5}$$

  $\mathscr{L}$ is the set of all classes, $c \in \mathscr{L}$ is a specific class, and TP_c, FP_c, FN_c are the true positives, false positives, and false negatives for class c, respectively.
- F1 Macro (F1_Macro): computes the F1 score independently for each class and then averages them, giving equal weight to every class regardless of its frequency.

$$F1_c = \frac{2 \cdot TP_c}{2 \cdot TP_c + FP_c + FN_c} \tag{6}$$

$$F1_{macro} = \frac{1}{|\mathscr{L}|} \sum_{c \in \mathscr{L}} F1_c \tag{7}$$

- Jaccard Similarity Index: measures the overlap between the predicted and the true label sets for each instance.

$$\text{Jaccard} = \frac{1}{N} \sum_{i=1}^{N} \frac{|Y_i \cap \hat{Y}_i|}{|Y_i \cup \hat{Y}_i|} \tag{8}$$

  where Y_i is the set of true labels and $\hat{Y}_i$ is the set of predicted labels for instance i, and N is the total number of instances.
- Set Accuracy: the strictest metric, requiring an exact match between the predicted label set and the true label set for each instance.

$$\text{Subset Accuracy} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\left(\hat{Y}_i = Y_i\right) \tag{9}$$

  where $\mathbf{1}(\cdot)$ is the indicator function, equal to 1 if the condition holds and 0 otherwise.

4.2. Experimental Setup

To evaluate the pipeline, a comparative approach was adopted with the CCDs generated by varying two key components of the procedure: the transcript used to produce the PR and the LLM prompt that extracts the CCD. The seven combinations considered are listed as follows: Basic Prompt + No Transcript, Basic Prompt + Random Transcript, Basic Prompt + Same Transcript, Our Prompt + No Transcript, Our Prompt + Least Related Transcript, Our Prompt + Random Transcript, Proposed Method.

The "Proposed Method" corresponds to the complete pipeline described in Section 3. The other methods consist of evaluating the LLM's ability to create a CCD using: a randomly chosen transcript (Random Transcript), the most (Same Transcript) and least (Least Related Transcript) related by cosine similarity, or no transcript at all (No Transcript). The methods are also divided between those that use the designated prompt with specific knowledge about CBT (Our prompt, reported in Section 3.2) and those with a baseline prompt (Basic Prompt) that defines some basic CBT guidelines but is not detailed.

Similarity measures were calculated using expert-generated CCDs from the PATIENT-Ψ-CM Cognitive Model Dataset, available upon request [13]. The dataset consists of 106 CCDs created by CBT experts using psychotherapy session summaries. These CCDs were used both as a gold standard for evaluation and as a reference for structuring CCDs; therefore, the fields of the diagrams are the same. The evaluation process involves using the Situation and History fields of the gold standard CCDs as Patient Description, which constitute the input of the pipeline. The generated CCD in JSON format is compared with the relevant gold standard, field by field, using the quantitative similarity metrics described in Section 4.1 and averaged between them. The Core Belief sets are excluded from this evaluation and are compared using the metrics related to the classifier evaluation, still using the Core Beliefs of the gold standard CCDs as a reference. The quality of the individual CCDs is assessed using an LLM-as-a-judge approach enriched with the quality definitions expressed in Section 4.1. The LLM is tasked with assigning one of the categories in the Ordinal Rating Scale, given as input the generated CCD and the initial Patient Description, to verify that the topics covered are consistent and relevant. The evaluations were conducted using 99 initial CCDs divided into 49 patients subjected to a specific situation, and 50 analogous patients subjected to a different situation. The remaining 7 CCDs from the original 106 were excluded, as they correspond to a third situational context not represented among the two patient typologies considered. The specific technologies used in the experiments are listed in the following. The summaries of the psychotherapy session transcripts and the creation of descriptions from the gold CCDs are generated using Mistral-7B-Instruction-v0.3[4]. The embedding model used in the RAG module and in the evaluation is sentence-transformers/all-mpnet-base-v2[5]. PR generation and CCD extraction are done with GPT-4o, while the LLM used for the CCD judgment is o1. All generations are done using temperature of 0.2 and top-p of 0.9. The pre-trained model on which the Core Beliefs classifiers are based is microsoft/deberta-v3-base[6]. In the evaluation process we used cross-encoder/stsb-roberta-base [7] as CrossEncoder STS model and cross-encoder/nli-deberta-v3-base [8] for the NLI Contradiction. To assess the quality of the CCDs, an LLM judge was employed using the following prompt, with minor paraphrasing for clarity:

> *You are an expert CBT supervisor.*
>
> *You will be given: (1) a PATIENT CASE VIGNETTE (case history + index situation); (2) a CBT case formulation purportedly describing that patient.*
>
> *Evaluate how well the CBT formulation explains the patient's presenting problems and maintaining mechanisms GIVEN the vignette.*
>
> *Quality is defined as a parsimonious, coherent, and meaningful CBT account of the client's presenting problems. A high-quality formulation links CBT elements (situations/triggers, ATs, emotions/physiology, behaviors, coping/compensatory strategies, underlying assumptions, core beliefs, and developmental/learning history) into plausible maintaining mechanisms across situations.*
>
> *The quality scale has four levels: Good, Good enough, Poor, Very poor:*
> *- Good enough: includes most relevant information (e.g., developmental factors, core beliefs, assumptions, coping/compensatory strategies) at an appropriate level of detail AND links it to prototypical problematic situations with at least a minimally coherent maintaining story consistent with the vignette.*
> *- Very poor: minimal integration of CBT elements, largely irrelevant content, or clear misunderstanding of CBT/CCD concepts.*

[4]https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3
[5]https://huggingface.co/sentence-transformers/all-mpnet-base-v2
[6]https://huggingface.co/microsoft/deberta-v3-base
[7]https://huggingface.co/cross-encoder/stsb-roberta-base
[8]https://huggingface.co/cross-encoder/nli-deberta-v3-base

- Good: stronger than "good enough" due to particularly insightful integration, clarity of maintaining mechanisms, or superior clinical utility that fits the vignette well.
- Poor: not "very poor," but still substantially limited (weak/unclear maintaining mechanisms, patchy links across levels, major gaps or mis-specified CBT elements).

IMPORTANT CONSTRAINTS:
- Do NOT evaluate formatting, diagram layout, headings, or whether the CCD structure is followed.
- The structure may be incomplete or unconventional: ignore this. Focus ONLY on the semantic content of the fields.
- Core beliefs can be abstract and not personalized; do not penalize abstraction when conceptually appropriate.
- Parsimony is a virtue: do not reward narrative richness unless it improves conceptual integration.
- Judge primarily: 1) functional integration among beliefs/assumptions/coping and maintaining situations; 2) clarity and plausibility of maintaining mechanisms; 3) coherence across levels (surface problems ↔ underlying beliefs/history); 4) consistency with the patient case vignette.

ANTI-BIAS:
- Do not default to "Good enough." Actively discriminate between adjacent levels.
- If mechanisms are clearly specified and intervention targets are obvious consider "Good."
- If links are vague, inconsistent or not grounded in the vignette consider "Poor" or "Very poor."

OUTPUT REQUIREMENTS:
- Assign exactly one rating: Good, Good enough, Poor, or Very poor.
- Focus ONLY on the semantic content of the fields; do NOT judge formatting or whether the CCD structure is perfectly followed.

PATIENT CASE VIGNETTE: {vignette}

CBT_FORMULATION: {cbt_text}

4.3. Results

The similarity with gold standard CCDs is calculated using four metrics commonly used in STS tasks: Embeddings Cosine Similarity (Emb. Cosine), CrossEncoder STS (CrossEnc STS), METEOR and NLI Contradiction (NLI Contr.), where, unlike the previous ones, the lower the value, the better. Table 3 shows the results obtained. Each value is the average over 99 CCDs, where each CCD score is itself the average across its fields. The Proposed Method achieves the best overall performance, outperforming all other configurations in METEOR and NLI Contradiction. On Embeddings Cosine Similarity, it ties with Our Prompt + Random Transcript, while on CrossEncoder STS it ties with both Our Prompt + Random Transcript and Our Prompt + No Transcript. The results show that the use of engineered prompt is the most influential variable, since each method using it produces CCDs closer to the gold standard. With regard to the use of transcriptions, the results show that using a transcript knowledge base is useful to add psychological depth to the CCD. The choice of the transcription most semantically related to the description is not a decisive factor in similarity, but it does show deterioration when a completely unrelated transcription is used.

The Core Beliefs classifiers are evaluated both using the inputs produced by the pipeline (i.e., Situation, AT, and IB LLM generated), the results of which are shown in Table 4, and using the same fields as the gold CCDs as inputs (using only Situation and ATs, since Intermediate Belief often cites the associated Core Belief), with the results shown in Table 5. In all methods using Our Prompt, Core Beliefs are predicted by the fine-tuned model, while the remaining approaches directly use the Core Beliefs hypothesized by the LLM. The tables show the metrics for both CBsM and CBsFG classification. As expected, the results show that the classification of the 3 CBsM is an easier task than the classification of the fine-grained ones. This is due to the number of labels (19 instead of 3), the much more nuanced concepts, and the smaller training dataset for CBsFG. The proposed classifier adopts a hierarchical approach consisting of two different models. The CBsM classification model is executed first, allowing the subsequent CBsFG classification model to activate only labels for which the relevant Major is active. As shown in Table 5, when evaluated on gold input fields, Our Classifier outperforms GPT-4o on most metrics, achieving notably higher scores on fine-grained labels ($F1_{\mu}$: 0.43 vs 0.32, Jaccard: 0.29 vs 0.22) and a modest

Method	Emb. Cosine	CrossEnc STS	METEOR	NLI Contr.
Basic Prompt + Same Transcript	0.42	0.37	0.10	0.14
Basic Prompt + Random Transcript	0.34	0.30	0.09	0.17
Basic Prompt + No Transcript	0.43	0.39	0.11	0.13
Our Prompt + Random Transcript	**0.52**	**0.45**	0.20	0.08
Our Prompt + Least Related Transcript	0.44	0.37	0.16	0.10
Our Prompt + No Transcript	0.49	**0.45**	0.16	0.08
Proposed Method	**0.52**	**0.45**	**0.21**	**0.06**

Table 3
Comparison of different configurations. Bold indicates the best result per metric. For NLI Contr., lower values are better; for all other metrics, higher values are better.

improvement on Major labels ($F1_{\mu}$: 0.78 vs 0.69). The only exception is $F1_{macro}$ CBsFG, where GPT-4o scores marginally higher (0.23 vs 0.22), suggesting a slight advantage in handling minority fine-grained classes.

Regarding the results on pipeline-generated inputs shown in Table 4, the Proposed Method achieves the best performance across nearly all metrics, with $F1_{\mu}$ CBsFG of 0.47 and $F1_{macro}$ CBsFG of 0.27, confirming that the combination of the engineered prompt and the semantically related transcript provides the most informative input for CBsFG classification. The only exceptions are $F1_{macro}$ CBsM, where Basic Prompt + Same Transcript ties with the Proposed Method (0.67), and Set Accuracy Major, where Basic Prompt + No Transcript achieves the highest score (0.59), suggesting that for exact label-set prediction on Majors, the absence of transcript noise can be beneficial. The Set Accuracy for fine-grained labels is 0.00 across all methods, which is expected given the large label space (19 classes), the multi-label nature of the task, and the strict exact-match requirement of this metric.

Method	$F1_{\mu}$ CBsM	$F1_{macro}$ CBsM	$F1_{\mu}$ CBsFG	$F1_{macro}$ CBsFG	Set Acc. CBsM	Jaccard CBsM	Jaccard CBsFG
Basic Prompt + Same Transcript	0.82	**0.67**	0.31	0.15	0.50	0.73	0.20
Basic Prompt + Random Transcript	0.79	0.60	0.28	0.14	0.40	0.68	0.18
Basic Prompt + No Transcript	0.82	0.60	0.33	0.18	**0.59**	0.74	0.23
Our Prompt + Random Transcript	0.80	0.63	0.45	0.22	0.49	0.71	0.30
Our Prompt + Least Related Transcript	0.76	0.53	0.42	0.20	0.41	0.67	0.28
Our Prompt + No Transcript	0.82	0.65	0.45	0.23	0.55	**0.75**	0.30
Proposed Method	**0.84**	**0.67**	**0.47**	**0.27**	0.51	**0.75**	**0.32**

Table 4
Classification performance across major and fine-grained CB labels. Best score per column is highlighted in bold. Set Acc. CBsFG is 0.00 for all methods.

Method	$F1_{\mu}$ CBsM	$F1_{macro}$ CBsM	$F1_{\mu}$ CBsFG	$F1_{macro}$ CBsFG	Set Acc. CBsM	Jaccard CBsM	Jaccard CBsFG
Our classifier	**0.78**	**0.62**	**0.43**	0.22	**0.45**	**0.69**	**0.29**
GPT-4o	0.69	0.59	0.32	**0.23**	0.32	0.60	0.22

Table 5
Classification results using *gold* expert-annotated input fields. The classifier is conditioned only on *situation* and *auto-thought*. Best score per column is highlighted in bold; Set Acc. CBsFG is 0.00 for both methods.

Qualitative assessment aims to evaluate whether a CCD is of good quality and correctly reflects the principles of CBT applied consistently to the patient description provided. To this aim, we used an LLM-as-a-judge approach, asking o1 (OpenAI) to evaluate the quality of our inferred CCDs according to the principles defined in [19]. Table 6 shows the LLM evaluation of the 99 CCDs produced using the seven combinations illustrated before. By assigning each label a decreasing numerical score (Good

= 3, Good Enough = 2, Poor = 1, Very Poor = 0), it was possible to calculate the Average Grade for each method. Furthermore, in order to obtain a robust overall ranking with respect to the distribution of preferences, the Borda Score was used, which allows ordinal ratings to be aggregated considering the entire distribution of judgments. The results clearly show that the proposed method achieves the best qualitative performance, with the highest Average Grade (2.51) and the highest Borda Score (248). These qualitative results are essentially consistent with the quantitative ones described before.

Method	Good (3)	Good Enough (2)	Poor (1)	Very Poor (0)	Average Grade	Borda Score
Basic Prompt + Same Transcript	0	39	60	0	1.39	138
Basic Prompt + No Transcript	0	28	71	0	1.28	127
Basic Prompt + Random Transcript	0	9	90	0	1.09	108
Our Prompt + Random Transcript	23	75	1	0	2.22	220
Our Prompt + Least Related Transcript	0	95	4	0	1.96	194
Our Prompt + No Transcript	0	58	41	0	1.59	157
Proposed Method	50	49	0	0	**2.51**	**248**

Table 6
Qualitative evaluation distribution with Average Grade and Borda Score. Ratings are from 3 (Good) to 0 (Very Poor).

Finally, to analyze the relationship between semantic similarity metrics and qualitative assessments of CCDs, high-quality (Good, Good Enough) and low-quality (Poor, Very Poor) groups were compared using Cliff's coefficient (Δ) as a non-parametric effect size measure. Figure 4 shows the Δ values obtained for each metric. METEOR shows the highest effect ($\Delta = 0.64$), followed by Embedding Cosine Similarity ($\Delta = 0.51$) and Cross-Encoder STS ($\Delta = 0.37$), indicating that CCDs rated as high quality systematically have higher similarity scores than those rated as low quality. The NLI Contradiction metric shows a small negative effect ($\Delta = -0.26$). Since in this case lower values indicate greater semantic consistency (lower probability of contradiction), the negative sign reflects an inverse relationship consistent with the expected interpretation: high-quality CCDs show lower probabilities of contradiction than low-quality ones. In probabilistic terms, the value of Δ can be interpreted as the likelihood that a randomly selected CCD from the high-quality group obtains a higher score than a randomly selected CCD from the low-quality group. This quantity is calculated as:

$$P(\text{High} > \text{Low}) = \frac{1 + \Delta}{2} \tag{10}$$

This means that there is approximately an 82% probability that a CCD belonging to the high-quality group will have a higher METEOR score than a CCD belonging to the low-quality group. This probabilistic interpretation makes the strength of the observed association particularly evident, indicating a marked separation between the two distributions. Overall, the results show a clear association between greater semantic similarity (and less contradiction) and higher quality of clinical formulation.

5. Conclusion and Future Work

The main goal of this work was to fill the gap caused by the lack of public datasets of CCD, which are models extensively used in CBT. To produce custom CCDs based on patients' profiles, we proposed a pipeline including machine learning classification, generative pretrained models, and a RAG technique to exploit real-world psychotherapy transcripts of reference cases. Since currently no formal semantic schema exists in the literature for representing CCD, we modeled a novel CCD ontology that describes causal chains between its fields. Our approach allowed the creation of a KG composed by 99 CCDs represented in RDF format and stored in a triple store. The triple store supports querying via SPARQL

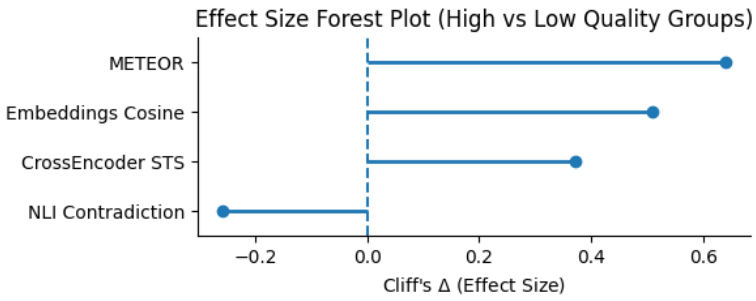

Figure 4: Forest plot of Cliff's Δ effect sizes comparing high- and low-quality groups across evaluation metrics. Positive values indicate higher scores for the high-quality group, while negative values indicate the opposite. Note that for NLI Contradiction, lower values correspond to better performance, so a negative Δ is the desirable direction for that metric.

language, providing a usable way to extract non-trivial relationships in the data. The evaluation conducted showed that our method allows the creation of CCDs that are semantically similar to those created by experts. Comparison with different baseline alternatives showed the advantages of using a detailed prompt including CBT knowledge. Results also indicate that the use of a psychotherapy session transcript from a similar case provides a useful knowledge base for refining the psychological aspects linked to reported facts.

Despite the promising achieved results, the current work has different limitations. First, our work lacks an experimental evaluation with human experts, which is needed to evaluate the effectiveness of our methods before integration into real-world applications. Moreover, CCD is often the result of an iterative process of variable duration between the psychotherapist and the patient. Therefore, it is important to highlight that the CCDs generated by our pipeline are initial hypotheses based on synthetic patient descriptions. In order to improve the accuracy of generated CCDs, future work will investigate an iterative generation mechanism including expert feedback to modify the CCD during the process and make it as suitable as possible to the specific case.

The proposed method is intended as a tool to complement and support established technologies in the field, rather than as a substitute for clinical practice. The formulation of a CCD is a delicate and iterative process that cannot be considered complete on the basis of a single transcript. The implementation of this system in real CBT sessions raises significant ethical concerns, particularly regarding patient privacy: providing sensitive clinical data to proprietary LLMs introduces risks of confidentiality breaches. The use of local models represents a partial mitigation, although it remains constrained by the limitations of the context window. From a clinical perspective, the summarization of therapeutic transcripts carries the risk of omitting subtle yet relevant details that distinguish one patient from another. Future work should explore more effective and clinically sensitive synthesis strategies. A further limitation lies in the use of semantically similar psychotherapy sessions: while proximity improves retrieval relevance, meaningful differences between cases may remain. In all intended applications, the clinician retains full responsibility for the final formulation, and the pipeline output should be treated as a preliminary hypothesis. Future work includes an empirical assessment of our dataset's usability from various perspectives, including its application in machine learning contexts.

Declaration on Generative AI

During the preparation of this work, the authors used **Google Gemini** in order to: refine the English language, correct grammar, and improve readability. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] S. Krokstad, D. A. Weiss, M. A. Krokstad, V. Rangul, K. Kvaløy, J. M. Ingul, O. Bjerkeset, J. Twenge, E. R. Sund, Divergent decennial trends in mental health according to age reveal poorer mental health for young people: repeated cross-sectional population-based surveys from the hunt study, norway, BMJ open 12 (2022) e057654.

[2] P. D. McGorry, C. Mei, A. Chanen, C. Hodges, M. Alvarez-Jimenez, E. Killackey, Designing and scaling up integrated youth mental health care, World Psychiatry 21 (2022) 61–76.

[3] J. M. Twenge, Why increases in adolescent depression may be linked to the technological environment, Current opinion in psychology 32 (2020) 89–94.

[4] M. R. Kader, M. M. Rahman, P. D. Bristi, F. Ahmmed, Change in mental health service utilization from pre-to post-covid-19 period in the united states, Heliyon 10 (2024).

[5] I. Starvaggi, L. Lorenzo-Luaces, et al., Psychotherapy access barriers and interest in digital mental health interventions among adults with treatment needs: Survey study, JMIR Mental Health 12 (2025) e65356.

[6] M. Casu, S. Triscari, S. Battiato, L. Guarnera, P. Caponnetto, Ai chatbots for mental health: a scoping review of effectiveness, feasibility, and applications, Applied Sciences 14 (2024) 5889.

[7] S. Harith, I. Backhaus, N. Mohbin, H. T. Ngo, S. Khoo, Effectiveness of digital mental health interventions for university students: an umbrella review, PeerJ 10 (2022) e13111.

[8] S. M. Massa, E. Porcu, D. Riboni, An initial investigation of mental well-being monitoring through personal healthcare devices, in: Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization, 2024, pp. 410–414.

[9] S. Zolfaghari, A. Kristoffersson, M. Folke, M. Lindén, D. Riboni, Unobtrusive cognitive assessment in smart-homes: Leveraging visual encoding and synthetic movement traces data mining, Sensors 24 (2024) 1381.

[10] X. Xu, B. Yao, Y. Dong, S. Gabriel, H. Yu, J. Hendler, M. Ghassemi, A. K. Dey, D. Wang, Mental-llm: Leveraging large language models for mental health prediction via online text data, Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 8 (2024) 1–32.

[11] S. Pinna, S. M. Massa, M. Fenu, G. Casti, D. Riboni, Integration of retrieval-augmented generation technique for llm-based differential diagnosis assistant, in: Proceedings of the 18th ACM International Conference on PErvasive Technologies Related to Assistive Environments, 2025, pp. 277–284.

[12] S. Pinna, S. M. Massa, D. Riboni, Ai-driven differential diagnosis: Leveraging rag and llms in intelligent healthcare systems, in: 2025 21st International Conference on Intelligent Environments (IE), IEEE, 2025, pp. 1–8.

[13] R. Wang, S. Milani, J. C. Chiu, J. Zhi, S. M. Eack, T. Labrum, S. M. Murphy, N. Jones, K. Hardy, H. Shen, et al., Patient-{\Psi}: Using large language models to simulate patients for training mental health professionals, arXiv preprint arXiv:2405.19660 (2024).

[14] H. Shen, Z. Li, M. Yang, M. Ni, Y. Tao, Z. Yu, W. Zheng, C. Xu, B. Hu, Are large language models possible to conduct cognitive behavioral therapy?, in: 2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE, 2024, pp. 3695–3700.

[15] B. A. Gaudiano, Cognitive-behavioural therapies: achievements and challenges, Evidence Based Mental Health 11 (2008).

[16] A. Mandal, T. Chakraborty, I. Gurevych, Towards privacy-aware mental health ai models, Nature Computational Science 5 (2025) 863–874.

[17] M. Jiang, Q. Zhao, J. Li, F. Wang, T. He, X. Cheng, B. X. Yang, G. W. Ho, G. Fu, A generic review of integrating artificial intelligence in cognitive behavioral therapy, arXiv preprint arXiv:2407.19422 (2024).

[18] J. S. Beck, Cognitive behavior therapy: Basics and beyond, Guilford Publications, 2020.

[19] W. Kuyken, C. D. Fothergill, M. Musa, P. Chadwick, The reliability and quality of cognitive case formulation, Behaviour research and therapy 43 (2005) 1187–1201.

[20] D. David, I. Cristea, S. G. Hofmann, Why cognitive behavioral therapy is the current gold standard

of psychotherapy, Frontiers in psychiatry 9 (2018) 4.
[21] J. B. Persons, The case formulation approach to cognitive-behavior therapy, Guilford Press, 2012.
[22] A. A. Abd-Alrazaq, M. Alajlani, A. A. Alalwan, B. M. Bewick, P. Gardner, M. Househ, An overview of the features of chatbots in mental health: A scoping review, International journal of medical informatics 132 (2019) 103978.
[23] N. Elsharawi, A. El Bolock, C-journal: A journaling application for detecting and classifying cognitive distortions using deep-learning based on a crowd-sourced dataset, in: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), 2024, pp. 3224–3234.
[24] M. Zhang, X. Yang, X. Zhang, T. Labrum, J. C. Chiu, S. M. Eack, F. Fang, W. Y. Wang, Z. Chen, Cbt-bench: Evaluating large language models on assisting cognitive behavior therapy, in: Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), 2025, pp. 3864–3900.
[25] H. Na, Cbt-llm: A chinese large language model for cognitive behavioral therapy-based mental health question answering, arXiv preprint arXiv:2403.16008 (2024).
[26] T. A. Furukawa, S. Iwata, M. Horikoshi, M. Sakata, R. Toyomoto, Y. Luo, A. Tajika, N. Kudo, E. Aramaki, Harnessing ai to optimize thought records and facilitate cognitive restructuring in smartphone cbt: An exploratory study, Cognitive Therapy and Research 47 (2023) 887–893.
[27] C. Peng, F. Xia, M. Naseriparsa, F. Osborne, Knowledge graphs: Opportunities and challenges, Artificial intelligence review 56 (2023) 13071–13102.
[28] B. Abu-Salih, M. Al-Qurishi, M. Alweshah, M. Al-Smadi, R. Alfayez, H. Saadeh, Healthcare knowledge graph construction: A systematic review of the state-of-the-art, open issues, and opportunities, Journal of Big Data 10 (2023) 81.
[29] L. Bos, K. Donnelly, Snomed-ct: The advanced terminology and coding system for ehealth, Stud Health Technol Inform 121 (2006) 279–290.
[30] L. M. Schriml, J. B. Munro, M. Schor, D. Olley, C. McCracken, V. Felix, J. A. Baron, R. Jackson, S. M. Bello, C. Bearer, et al., The human disease ontology 2022 update, Nucleic acids research 50 (2022) D1255–D1261.
[31] G. Fragoso, S. de Coronado, M. Haber, F. Hartel, L. Wright, Overview and utilization of the nci thesaurus, Comparative and functional genomics 5 (2004) 648–654.
[32] M. Rotmensch, Y. Halpern, A. Tlimat, S. Horng, D. Sontag, Learning a health knowledge graph from electronic medical records, Scientific reports 7 (2017) 5994.
[33] M. G. McInnis, B. Coleman, E. Hurwitz, P. N. Robinson, A. E. Williams, M. A. Haendel, J. A. McMurry, Integrating knowledge: The power of ontologies in psychiatric research and clinical informatics., Biological Psychiatry (2025).
[34] S. Freidel, E. Schwarz, Knowledge graphs in psychiatric research: Potential applications and future perspectives, Acta Psychiatrica Scandinavica 151 (2025) 180–191.
[35] S. Gao, K. Yu, Y. Yang, S. Yu, C. Shi, X. Wang, N. Tang, H. Zhu, Large language model powered knowledge graph construction for mental health exploration, Nature Communications 16 (2025) 7526.
[36] X. Liu, X. Zheng, W. Cui, Psychological counseling with integration of knowledge graph and multi-agent collaboration, Journal of Shanghai Jiaotong University (Science) (2024) 1–13.
[37] A. Nair, S. S. R. Abidi, W. Van Woensel, S. Abidi, Ontology-based personalized cognitive behavioural plans for patients with mild depression, in: Public Health and Informatics, IOS Press, 2021, pp. 729–733.
[38] L. M. R. Barahona, B.-H. Tseng, Y. Dai, C. Mansfield, O. Ramadan, S. Ultes, M. Crawford, M. Gasic, Deep learning for language understanding of mental health concepts derived from cognitive behavioural therapy, in: Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis, 2018, pp. 44–54.
[39] M. Kim, D. Yoo, Y. Hwang, M. Kang, N. Kim, M. Gwak, B.-w. Kwak, H. Chae, H. Kim, Y. Lee, et al., Can you share your story? modeling clients' metacognition and openness for llm therapist evaluation, in: Findings of the Association for Computational Linguistics: ACL 2025, 2025, pp.

25943–25962.
[40] D. Cer, M. Diab, E. Agirre, I. Lopez-Gazpio, L. Specia, Semeval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation, in: Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017), 2017, pp. 1–14.
[41] H. Steck, C. Ekanadham, N. Kallus, Is cosine-similarity of embeddings really about similarity?, in: Companion Proceedings of the ACM Web Conference 2024, 2024, pp. 887–890.
[42] N. Thakur, N. Reimers, J. Daxenberger, I. Gurevych, Augmented sbert: Data augmentation method for improving bi-encoders for pairwise sentence scoring tasks, in: Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: Human language technologies, 2021, pp. 296–310.
[43] A. Lavie, M. J. Denkowski, The meteor metric for automatic evaluation of machine translation, Machine translation 23 (2009) 105–115.
[44] O.-M. Camburu, T. Rocktäschel, T. Lukasiewicz, P. Blunsom, e-snli: Natural language inference with natural language explanations, Advances in Neural Information Processing Systems 31 (2018).
[45] K. Meissel, E. S. Yao, Using cliff's delta as a non-parametric effect size measure: an accessible web app and r tutorial, Practical Assessment, Research, and Evaluation 29 (2024).