Intégrer Récits cliniques et Questionnaires Likert : une approche fondée sur des dialogues patient–médecin simulés

Cyrielle Mogoun[1] Frederic Bechet[1] Magalie Ochs[1] Laurent Boyer[2]
(1) LIS CNRS, Aix Marseille Université
(2) CEReSS, Aix-Marseille Université
nom.prenom@lis-lab.fr

RÉSUMÉ

L'évaluation des patients en santé mentale repose à la fois sur des conversations ouvertes et des questionnaires structurés avec des échelles de type Likert. La correspondance entre des conversations ouvertes et des réponses aux questionnaires reste aujourd'hui un défi. Cet article explore l'utilisation du Traitement Automatique du Langage naturel (TAL) pour établir une correspondance entre les dialogues patient-médecin et des réponses standardisées à des questionnaires. Afin de pallier le manque de données annotées, nous proposons un cadre de génération de données synthétiques fondé sur des conversations simulées associées à des scores Likert d'un questionnaire de santé mentale. En exploitant des modèles génératifs pour produire des dialogues cliniques réalistes associés à des annotations standardisées, nous visons à produire des ressources permettant d'entraîner des modèles robustes capables de traduire directement des transcriptions conversationnelles en réponses structurées à des questionnaires.

ABSTRACT

The assessment of mental health patients relies on both open-ended conversations and structured questionnaires using Likert-type scales. Matching open-ended conversations with questionnaire responses remains a challenge today. This article explores the use of Natural Language Processing (NLP) to establish a correspondence between patient-physician dialogues and standardized questionnaire responses. To address the lack of annotated data, we propose a synthetic data generation framework based on simulated conversations paired with Likert scores from a mental health questionnaire. By leveraging generative models to produce realistic clinical dialogues paired with standardized annotations, we aim to create resources for training robust models capable of directly translating conversational transcripts into structured questionnaire responses.

MOTS-CLÉS : Traitement automatique du langage naturel (TALN), Dialogues cliniques patient–médecin, Échelles de Likert, Données synthétiques, Conversion conversation–questionnaire.

KEYWORDS: Automatic Natural Language Processing (NLP), Patient–doctors dialogues, Likert scales, Synthetic data, Conversation–questionnaire conversion.

1 Introduction

En recherche médicale, l'évaluation des patients repose généralement sur deux paradigmes complémentaires : des interactions qualitatives ouvertes, telles que les conversations patient–médecin, et des instruments quantitatifs structurés, tels que les questionnaires à échelle de Likert. Si le dialogue libre permet de capturer des informations riches, contextuelles et subjectives, les questionnaires Likert permettent une mesure standardisée et des analyses statistiques à grande échelle. L'articulation entre ces modalités demeure un défi central, en particulier dans les contextes cliniques où l'interprétabilité et le passage à l'échelle sont essentiels.

Les avancées récentes en traitement automatique du langage naturel (TALN) ont permis l'extraction automatique d'informations structurées à partir de textes cliniques non structurés. En particulier, des travaux antérieurs ont exploré la faisabilité de la mise en correspondance entre des données conversationnelles et des réponses à des questionnaires, notamment via des approches zéro-shot pour le remplissage de questionnaires cliniques à partir d'interactions humain–machine (Toudeshki *et al.*, 2021). Tout en soulignant la difficulté de la tâche qui reste aujourd'hui non résolue, ces approches mettent en évidence le potentiel de l'analyse automatique de questionnaire pour prédire les expressions qualitatives des patients en métriques cliniques quantitatives.

Cependant, une limitation majeure dans ce domaine réside dans la rareté de jeux de données conversationnels de haute qualité, anonymisés, reliant les dialogues patients à des résultats de questionnaires validés. Pour répondre à ce manque, cet article propose l'utilisation de données synthétiques générées à partir de conversations simulées entre patients et médecins. En exploitant des modèles génératifs pour produire des dialogues cliniques réalistes associés à des annotations standardisées sur échelle de Likert, nous visons à entraîner des modèles robustes capables de traduire directement des transcriptions conversationnelles en réponses structurées à des questionnaires.

Ce travail contribue ainsi à l'intersection émergente entre TALN médical et évaluation mixte en (i) étudiant la relation entre modalités d'évaluation qualitatives et quantitatives, et (ii) introduisant un cadre de génération de données visant à surmonter les limitations actuelles de l'apprentissage supervisé pour la mise en correspondance entre conversations et questionnaires.

2 État de l'art

L'intégration des interfaces conversationnelles en santé a fait l'objet de nombreux travaux, notamment dans le contexte de l'engagement des patients et de la collecte de données. Les agents conversationnels se sont révélés constituer des alternatives efficaces aux questionnaires traditionnels, en améliorant l'expérience utilisateur et la qualité des réponses (Pas *et al.*, 2020). De même, des systèmes ont été déployés dans des contextes cliniques tels que le suivi post-partum, démontrant la faisabilité d'extraire des informations pertinentes à partir des interactions avec les patients (Leitner *et al.*, 2025).

D'un point de vue méthodologique, des travaux antérieurs ont exploré des approches automatisées de complétion de questionnaires à partir d'entrées conversationnelles. En particulier, Toudeshki *et al.* (Toudeshki *et al.*, 2021) ont introduit un cadre zéro-shot pour le remplissage de questionnaires cliniques, illustrant comment des données issues de dialogues peuvent être mises en correspondance avec des sorties structurées sans apprentissage spécifique à la tâche. Parallèlement, des études récentes ont examiné le rôle des grands modèles de langage dans la conception et la génération de ques-

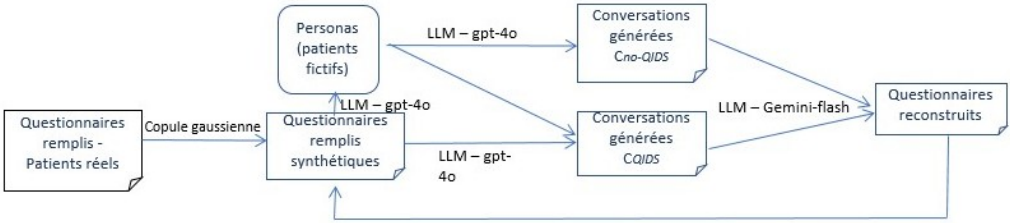

FIGURE 1 – Pipeline pour la génération de données et leur évaluation

tionnaires médicaux, soulignant davantage l'interaction entre langage naturel et outils d'évaluation structurés (Coraci *et al.*, 2023).

Malgré ces avancées, les travaux existants reposent en grande partie sur des jeux de données limités ou propriétaires, ce qui restreint le développement des méthodes proposées. En outre, la modélisation explicite de la transition entre données conversationnelles qualitatives et réponses quantitatives sur échelle de Likert reste encore peu explorée. Ce travail s'inscrit dans la continuité de ces efforts en combinant des techniques d'analyse de documents avec la génération de données synthétiques (Tayal *et al.*, 2025), permettant un entraînement de modèles unifiant ces deux paradigmes d'évaluation.

3 Méthodologie

L'objectif de ce travail est de remplacer les évaluations traditionnelles de la qualité de vie fondées sur des questionnaires par une extraction automatisée d'informations structurées à partir de conversations entre patient et agent conversationnel. Plus précisément, étant donné un dialogue entre un patient et un agent conversationnel, le but est d'inférer les réponses correspondantes à un questionnaire sur échelle de Likert.

Une contrainte majeure réside dans l'interdiction d'utiliser des données conversationnelles réelles de patients avec des grands modèles de langage (LLM) commerciaux accessibles via API, en raison de considérations de confidentialité et de réglementation. De plus, il n'existe pas de jeu de données « écologique » à grande échelle alignant conversations et questionnaires complétés, ce qui rend l'apprentissage supervisé difficile à partir de seules données réelles.

Pour répondre à ces limitations, nous proposons, dans cet article, un pipeline de génération de données synthétiques permettant une distillation de connaissances depuis des LLM commerciaux de grande taille vers un modèle plus compact, déployable localement. L'approche consiste à générer des données de questionnaires artificielles mais statistiquement plausibles, à les enrichir avec des personas de patients, puis à produire des conversations correspondantes cohérentes avec les réponses aux questionnaires. La Figure 1 illustre l'approche méthodologique présentée dans ce papier. A partir d'un ensemble de questionnaires remplis par des patients réels, une méthode fondée sur une copule gaussienne est utilisée pour générer des données équivalentes synthétiques (Section 4.2). Ces données combinées avec des personas de patients fictifs (Section 4.3) sont utilisées pour générer des conversations médecins-patients à partir d'un LLM (Section 4.4). En utilisant un autre LLM, la capacité de recontruction des réponses aux questionnaires à partir des conversations est évaluée

(Section 4.5 et 4.6).

L'objectif est de pouvoir créer un jeu de données synthétique constitué de triplets alignés (questionnaire, persona, conversation), pour, à terme, entraîner un LLM compact capable de réaliser à la fois l'interaction conversationnelle et l'inférence des réponses au questionnaire.

4 Expériences

Le cadre applicatif choisi s'inscrit dans le domaine des études en santé mentale, plus précisément pour l'évaluation de la dépression à partir de la situation de santé auto-déclarée à l'aide du questionnaire *Quick Inventory of Depressive Symptomatology (QIDS)* (Rush *et al.*, 2003). Ce questionnaire est un outil standardisé développé comme une version concise d'échelles d'évaluation de la dépression antérieures introduites en 1986, afin de garantir une évaluation fiable et reproductible des épisodes dépressifs majeurs. Dérivé de l'*Inventory of Depressive Symptomatology (IDS)* comportant 30 items, qui existe à la fois en version évaluée par un clinicien et en auto-évaluation, le QIDS comprend 16 items et est également disponible en deux versions. La version auto-administrée, QIDS-SR16, est conçue pour être remplie par le patient et fournit une évaluation simplifiée mais robuste de la sévérité des symptômes dépressifs. Chaque symptôme est évalué sur une échelle uniforme de 0 à 3, avec des critères clairement définis de fréquence et de sévérité. Le questionnaire couvre les neuf domaines diagnostiques fondamentaux de la dépression majeure tels que définis par les critères du DSM — tels que l'humeur, le sommeil, l'appétit, l'énergie, la concentration et les idées suicidaires — et génère un score total allant de 0 à 27, reflétant la sévérité actuelle des symptômes dépressifs au cours des sept derniers jours, indépendamment de leur durée ou de leur apparition.

4.1 Prétraitement et représentation des données

Pour mettre en œuvre notre méthodologie, nous sommes partis d'un jeu de données composé de 164 questionnaires remplis par de vrais patients, anonymisés (les seules informations démographiques disponibles étant le genre et l'âge). Pour chaque question, la réponse est une valeur numérique sur une échelle de Likert allant de 0 à 3. Nous avons sélectionné dans cet ensemble les questionnaires de patients souffrant de troubles dépressifs et ayant répondu à toutes les questions non exclusives, excluant ainsi un total de 40 patients. Notre jeu de données comprend donc 124 patients dont le diagnostic correspond à l'une des 5 catégories obtenu à partir d'un score total calculé à partir des scores associés à chaque question : normal / absence de dépression (0-5), dépression légère (6-10), dépression modérée (11-15), dépression sévère (16-20) et dépression très sévère (21-27). L'âge a été calculé et normalisé à l'aide de la formulation suivante $Age = (date\_evaluation - date\_naissance)/365.25$ et le genre (femme/homme) a été encodé numériquement afin de permettre son intégration dans le cadre de la modélisation statistique.

Un défi majeur provient de la structure du questionnaire QIDS-SR16 : plusieurs paires d'items sont mutuellement exclusives (par exemple, *augmentation vs diminution de l'appétit*, *augmentation vs diminution du poids*). En conséquence, le jeu de données contient des valeurs manquantes structurelles, où un seul item de chaque paire est renseigné étant donnée la conception du questionnaire.

Pour traiter ce problème, nous avons adopté une stratégie en deux étapes :

— Les valeurs manquantes non structurelles (c'est-à-dire non liées à la conception du question-

naire) ont été supprimées afin de garantir l'intégrité des données et éviter l'introduction de bruit dans le processus d'apprentissage de la distribution.
— Les valeurs manquantes structurelles (résultant d'items mutuellement exclusifs) ont été conservées et encodées à l'aide d'une valeur sentinelle (-1). Cet encodage représente explicitement l'absence de réponse due à la logique du questionnaire, sans introduire d'information artificielle ni fusionner des dimensions cliniques distinctes.

Contrairement aux approches qui fusionnent des variables mutuellement exclusives en une seule caractéristique, cette représentation préserve la structure originale du questionnaire et évite d'imposer des hypothèses de symétrie ou d'équivalence entre des symptômes opposés.

Ce choix de conception est soutenu par des travaux antérieurs soulignant qu'un traitement inapproprié des données manquantes — en particulier lorsqu'elles sont induites structurellement — peut conduire à des estimations biaisées des paramètres et à des structures de dépendance déformées dans les modèles statistiques (Little & Rubin, 2002). En encodant explicitement l'absence structurelle, nous préservons à la fois l'interprétabilité et la validité statistique du jeu de données.

4.2 Modélisation de la distribution du questionnaire

Pour modéliser la distribution jointe des réponses au questionnaire et des variables démographiques, nous avons implémenté une approche fondée sur une copule gaussienne adaptée aux données mixtes et discrètes. Cette méthodologie s'inspire de travaux récents sur la génération de données synthétiques par copules, qui soulignent l'importance de combiner une estimation empirique des marginales avec une modélisation robuste des dépendances pour les jeux de données tabulaires (Houssou *et al.*, 2022).

Les réponses au QIDS-SR16, initialement exprimées sur une échelle de Likert, ont été traitées comme des variables catégorielles. Afin d'appliquer le cadre des copules, ces variables ont été transformées en représentations numériques tout en préservant leurs distributions empiriques. Cette transformation repose sur une approche fondée sur les rangs, où les catégories les plus fréquentes sont associées à des probabilités cumulées plus élevées, permettant une projection cohérente dans un espace continu. En revanche, les variables démographiques, telles que l'âge et le genre, ont été conservées sous leur forme numérique et directement intégrées dans le processus de modélisation.

La procédure globale suit trois étapes principales :

1. Estimation des marginales : des fonctions de répartition empirique (ECDF) sont calculées pour chaque variable, permettant une estimation non paramétrique des distributions marginales.
2. Modélisation des dépendances : les variables sont projetées dans un espace gaussien latent, où les dépendances sont estimées à l'aide de mesures de corrélation fondées sur les rangs (tau de Kendall), plus robustes pour des données non normales et ordinales.
3. Échantillonnage et reconstruction : de nouveaux échantillons sont générés dans l'espace latent puis projetés dans l'espace des données originales via des transformations inverses, garantissant la préservation à la fois des marginales et des dépendances.

Cette approche permet de générer des questionnaires synthétiques qui restent statistiquement cohérents avec le jeu de données original tout en prenant en compte la nature discrète des variables cliniques.

En utilisant la distribution jointe apprise, nous avons généré un ensemble de 124 questionnaires de patients entièrement synthétiques, correspondant à la taille du jeu de données original. Ce choix de

conception permet une comparaison contrôlée entre données réelles et synthétiques tout en conservant l'échelle initiale des données. Dans la suite de cette étude, nous utiliserons uniquement ce nouveau jeu de données généré, car il peut être partagé et utilisé dans des études de reproductibilité puisqu'il ne correspond pas directement à des patients réels mais à des patients simulés. Les questionnaires générés ont été évalués afin de vérifier qu'ils reproduisent fidèlement les propriétés statistiques du jeu de données original.

4.3 Génération de personas

Pour chaque questionnaire synthétique, trois personas distincts (c.a.d. une représentation fictive d'un patient) ont été générés en anglais, ce qui donne un total de 372 personas (124×3) . Cette multiplicité permet d'introduire de la variabilité dans les représentations des patients tout en maintenant la cohérence avec un même profil clinique sous-jacent.

Les personas ont été générés à l'aide d'un modèle de langage de grande taille (gpt4.O) via une ingénierie de prompts soigneusement conçue. Les prompts instruisaient explicitement le modèle de produire des descriptions de patients cohérentes et réalistes, alignées avec les réponses correspondantes au questionnaire. Cette approche garantit que chaque persona reste compatible avec le niveau de sévérité des symptômes encodé dans le questionnaire synthétique.

Chaque persona généré inclut des informations démographiques (par exemple, âge, genre), des caractéristiques psychologiques et comportementales, ainsi que le contexte social et de vie (par exemple, situation familiale, profession, habitudes quotidiennes).

En s'inspirant des travaux proposés dans Jiang *et al.* (Jiang *et al.*, 2024) pour la génération d'histoires, afin de structurer davantage et de diversifier les personas générés, nous avons intégré des traits de personnalité fondés sur le modèle des Big Five (Ouverture, Conscienciosité, Extraversion, Agréabilité, Névrosisme). Le modèle des Big Five est un cadre largement reconnu en psychologie pour décrire la personnalité humaine selon cinq dimensions continues (McCrae & John, 1992). Ces traits capturent des tendances comportementales stables et sont particulièrement utiles pour modéliser l'hétérogénéité des réponses humaines.

Dans notre pipeline, plutôt que d'imposer un modèle rigide, la stratégie de prompting encourage une variabilité contrôlée, permettant au modèle de produire des profils diversifiés mais plausibles. Nous avons également utilisé le prompting pour inférer et intégrer implicitement les traits des Big Five dans chaque description de persona. Cela remplit deux objectifs principaux :

- Amélioration du réalisme : les traits de personnalité influencent la manière dont les individus expriment leurs symptômes, émotions et comportements en conversation ;
- Augmentation de la diversité : plusieurs personas générés à partir d'un même questionnaire peuvent différer par leurs profils de personnalité, conduisant à des styles conversationnels variés malgré des scores cliniques identiques.

Par exemple, deux individus présentant une sévérité de dépression similaire peuvent différer significativement dans la manière dont ils décrivent leurs expériences en fonction de leur niveau d'extraversion ou de névrosisme.

Des travaux récents ont montré que les LLMs peuvent simuler efficacement des agents de type humain présentant des comportements cohérents lorsqu'ils sont guidés par des prompts structurés et des contraintes contextuelles (Gao *et al.*, 2023). Dans ce contexte, la combinaison de la cohérence clinique (questionnaire) et de la variabilité psychologique (Big Five) peut aider à générer des agents

conversationnels réalistes. La table 1 présente un exemple de persona généré illustrant l'ensemble des caractéristiques d'un patient simulé.

```
-- demographics : age:55 , sex:F
-- life-context
    -- occupation: Accountant
    -- family-situation: Married, no children
    -- living-environment: Urban apartment
    -- stressors: High-pressure job, financial insecurity due to partner's
       unemployment
    -- social-support: Limited, primarily relies on spouse
-- clinical-profile
    -- sleep: Difficulty sleeping despite fatigue
    -- mood: Frequently irritable and anxious
    -- appetite: Unchanged
    -- cognition: Occasional forgetfulness
    -- self-perception: Critical of self, feels burdened by financial stress
    -- suicidal-ideation: Denies
    -- anhedonia: Complete, no interest in formerly enjoyable activities
    -- energy: Low, fatigue is persistent
    -- psychomotor: Lethargic
-- personality
    -- openness: 0.5
    -- conscientiousness: 0.9
    -- extraversion: 0.3
    -- agreeableness: 0.5,
    -- neuroticism: 0.8
-- narrative-summary
    -- This 55-year-old female works as an accountant dealing with significant stress
       from her job and financial pressures due to her partner's unemployment. She's
       irritable and anxious, with low energy and difficulty finding joy in activities
       she once enjoyed. Her emotional resilience is challenged by limited social
       support, as she primarily relies on her spouse for support
    -- inference-notes: Financial insecurity and high job pressure significantly
       affect her mood and energy levels, impacting her ability to enjoy life and
       increasing irritability.
```

TABLE 1 – Exemple de persona généré

4.4 Génération de dialogues sous conditions contrôlées

Tous les dialogues ont été générés en anglais à l'aide d'un modèle de langage de grande taille (LLM) accessible via une API (*gpt4.o*), en s'appuyant sur des stratégies de prompting soigneusement conçues. Les prompts ont été élaborés afin d'assurer à la fois une cohérence clinique avec le questionnaire sous-jacent et un réalisme comportemental dérivé des caractéristiques des personas.

Un aspect clé de notre approche est que la génération de dialogues ne repose pas uniquement sur le contenu des personas, mais également sur des traits de personnalité issus du cadre des Big Five, explicitement intégrés dans le processus de prompting. Cela permet au modèle de générer des conversations qui diffèrent non seulement par leur contenu, mais aussi par leur style, leur ton et leur expressivité (Jiang *et al.*, 2024).

Afin d'étudier l'impact d'un conditionnement structuré sur les données conversationnelles, deux stratégies de génération de dialogues ont été mises en œuvre, aboutissant à deux corpus distincts :

— **Condition persona seule (Corpus $C_{\texttt{no-QIDS}}$)** : dans la première condition, les dialogues sont générés en utilisant uniquement la description du persona comme entrée. Notre hypothèse est que cette configuration favorise des conversations naturelles et peu contraintes, le modèle

devant exprimer implicitement l'état du patient à travers le dialogue sans accès direct aux données structurées du questionnaire.

— **Condition persona + questionnaire (Corpus $C_{\texttt{QIDS}}$)** : dans la seconde condition, les dialogues sont générés en utilisant à la fois le persona et le questionnaire associé (fourni au format JSON structuré). Ce conditionnement supplémentaire devrait contraindre le modèle à intégrer des informations cliniques spécifiques pouvant aider à la reconstruction du questionnaire. Cependant ces contraintes peuvent nuire au naturel des conversations générées.

Pour chacun des 372 personas, deux dialogues ont été générés, un dans chacune des conditions, aboutissant à un total de 744 dialogues, constituant deux corpus parallèles (de 372 dialogues) pour une comparaison contrôlée.

4.5 Évaluation de la qualité des conversations générées

Les conversations générées par notre méthode avec les deux conditions (avec et sans la présence du questionnaire) peuvent être évaluées selon deux dimensions :

— *plausibilité* : est-ce que les conversations générées sont réalistes ? ressemblent-elles à de *vraies* interactions patients/docteur ?

— *pertinence* : est-ce que les conversations permettent d'inférer les mêmes réponses aux questionnaires que ceux qui ont servis à les générer ?

Idéalement il faudrait mener une enquête auprès de médecins ayant l'expérience de mener ce type d'entretiens pour avoir leur retour. Cette tâche est prévue dans le cadre de notre projet, mais dans un premier temps nous avons mené une première expérience implémentant la méthode "*LLM-as-a-judge*" pour juger de la plausibilité et la pertinence des conversations générées.

À cet effet, chaque conversation générée a été traité par un second LLM (Gemini), à travers deux tâches : évaluer l'aspect naturel de chaque dialogue sur une échelle de 0 à 10 grâce à un prompt simple et explicite ; reconstruire le questionnaire QIDS-SR16 au format JSON à partir de la transcription de la conversation avec un prompt structurant les réponses selon le thème des questions.

Cette approche d'évaluation d'un LLM par un autre LLM s'est imposée comme une alternative acceptable à l'évaluation humaine dans des tâches de traitement automatique pour lesquels il n'y a pas de corpus de référence sur lesquels se comparer. En particulier, les travaux de Jiang *et al.*(Jiang *et al.*, 2024) montrent la capacité des LLMs a juger correctement l'aspect naturel des textes générés en comparaison à des annotation humaines.

Les résultats présentés dans la table 2 sur l'évaluation de l'aspect naturel montrent une différence entre les deux conditions. Les dialogues générés sans conditionnement par le questionnaire ($C_{\texttt{no-QIDS}}$) présentent des scores sur l'aspect naturel des dialogues significativement plus élevés. Les écarts-types relativement faibles dans les deux conditions indiquent que ces résultats sont stables à travers les échantillons et ne sont pas dus à des valeurs aberrantes.

Ceci conforte notre hypothèse que les dialogues générés uniquement à partir de personas semblent donc plus naturels que ceux contraints par des informations issues de questionnaires structurés.

Afin d'évaluer la capacité du modèle à reconstruire des informations cliniques à partir des dialogues, les questionnaires reconstruits ont été comparés aux questionnaires synthétiques originaux à l'aide de métriques classiques de classification (Exactitude, Précision, Rappel, Score F1 et Score F1 pondéré). Ces métriques ont été calculées pour les cinq niveaux de sévérité de la dépression dérivés des scores

QIDS-SR16 (normale, lègère, modéré, sévère, très sévère). Les résultats sont reportés dans les tables 2 et 3.

C	Exact.	F1	F1 pondéré	Préc.	Rappel	Naturel
$C_{\texttt{QIDS}}$	0.65	0.48	0.65	0.49	0.55	6.47 (ET=0.67)
$C_{\texttt{no-QIDS}}$	0.44	0.28	0.46	0.27	0.45	8.60 (ET=0.56)

TABLE 2 – Comparaison des corpus de test $C_{\texttt{QIDS}}$ et $C_{\texttt{no-QIDS}}$ en termes de naturel des conversations (score entre 0 et 10) et de la capacité d'extraction des réponses aux questionnaires à partir des conversations.

	$C_{\texttt{no-QIDS}}$			$C_{\texttt{QIDS}}$			
diag./metriques	**Préc.**	**Rappel**	**F1**	**Préc.**	**Rappel**	**F1**	**Nb. dialogues**
Normale	0.09	1.00	0.16	0.33	0.67	0.44	3
Légère	0.20	0.39	0.27	0.51	0.69	0.58	51
Modérée	0.67	0.51	0.58	0.80	0.72	0.76	222
Sévère	0.41	0.37	0.39	0.48	0.61	0.54	75
Très sévère	0.00	0.00	0.00	0.33	0.05	0.08	21

TABLE 3 – Performances d'extraction des réponses aux questionnaires à partir du corpus $C_{\texttt{no-QIDS}}$ et $C_{\texttt{QIDS}}$

Lorsque les dialogues sont générés avec accès au questionnaire (corpus $C_{\texttt{QIDS}}$), l'exactitude atteint 0.65 et le F1-score pondéré est de 0.65 (2). Comme montré dans les résultats reportés dans le tableau 3, la catégorie de dépression *modérée* est bien capturée (F1-score = 0.76), avec une majorité de cas correctement classés (159 / 222) ; en revanche, les catégories *légère* et *sévère* présentent des performances modérées (F1 = 0.58 pour *légère*, 0.54 pour *sévère*). Les catégories extrêmes (*normale*, *très sévère*) sont mal capturées (rappel *très sévère* = 0.05).

Ces résultats montrent que les conversations générées concernent principalement les catégories cliniques centrales, qui dominent le jeu de données, mais correspondent mal aux cas extrêmes, ce qui n'est pas surprenant étant donné la tendance des LLM à *lisser* le texte généré.

Lorsque les dialogues sont générés sans conditionnement par le questionnaire (corpus $C_{\texttt{no-QIDS}}$), on observe une baisse d'exactitude de 0.21, avec un F1-score pondéré de 0.46 (table 2). Comme montré dans les résultats reportés dans la table 3, la dégradation est significative dans toutes les catégories. La catégorie *modérée* reste partiellement identifiable (F1-score = 0.58 contre 0.76 avec QIDS), tandis que la catégorie *très sévère* disparaît complètement (précision = 0.00, rappel = 0.00). On observe également une augmentation marquée des erreurs de classification vers les catégories basses ou intermédiaires. Sans guidage structuré, le modèle ne parvient pas à encoder de manière fiable les signaux cliniques nécessaires dans le dialogue, ce qui entraîne une perte d'information et un effondrement des catégories.

La comparaison entre les deux conditions révèle un écart de performance substantiel. L'utilisation du questionnaire entraîne des gains significatifs en rappel pour les catégories *modérée* et *sévère*, ainsi qu'une meilleure couverture globale des classes (moins de prédictions manquantes). Avec le questionnaire QIDS, les erreurs restent majoritairement locales (par exemple entre *modérée* et *sévère*), tandis que sans le questionnaire QIDS, elles deviennent plus globales et désorganisées, avec des confusions fréquentes vers les catégories *légère/modérée* et la disparition des classes rares.

Ces résultats soutiennent fortement les hypothèses initiales selon lesquelles la condition Persona + Questionnaire QIDS permet de produire des dialogues plus contrôlés et d'améliorer significativement la précision de reconstruction ainsi que la fiabilité diagnostique.

5 Conclusion

Dans cet article, nous avons proposé un pipeline entièrement synthétique de génération de données, conçu pour surmonter les contraintes de confidentialité inhérentes aux données cliniques. À partir d'un jeu de données limité de 124 questionnaires QIDS-SR16, nous avons modélisé la distribution jointe des réponses à l'aide d'une approche par copule gaussienne, permettant la génération de patients synthétiques statistiquement cohérents. Ces profils synthétiques ont ensuite été enrichis par une génération de personas fondée sur les LLM, intégrant des traits démographiques, contextuels et psychologiques, notamment via le modèle issu de la psychologie du Big Five. Enfin, nous avons généré deux corpus distincts de dialogues clinicien–patient à l'aide de stratégies de prompting : l'un fondé uniquement sur les personas, et l'autre conditionné à la fois par les personas et les données structurées du questionnaire. Les données ont été générées en anglais mais la pipeline proposée pourrait être adaptée aisément à d'autres langues.

Nos résultats expérimentaux mettent en évidence un compromis. D'un côté, Les conversations patient-médecin générées à partir de personas de patients fictifs construits sur la base de questionnaires préremplis sont significativement plus naturelles que lorsque la génération est contrainte en plus par les réponses aux questionnaires. Cependant, ces conversations ne permettent pas d'encoder de manière fiable l'information clinique, conduisant à des performances limitées de reconstruction des résultats des questionnaires. D'un autre côté, les conversations générées à partir du persona du patient et des réponses aux questionnaires permettent d'améliorer significativement l'exactitude de reconstruction des réponses aux questionnaires, confirmant que le conditionnement structuré renforce l'extractabilité des signaux cliniques, au prix toutefois de conversations moins naturelles.

Pour évaluer l'impact de la qualité des modèles sur la génération, nous avons testé sur un corpus de 8 personas avec des diagnostics différents, la performance du modèle *gpt-5-1* sous conditionnement QIDS. Les résultats montrent que ce modèle apporte un compromis permettant d'améliorer l'aspect naturel des conversations générées (M=7, E.T=0.93) comparativement au modèle *gpt-40* avec QIDS mais une légère baisse des performances pour la reconstruction des questionnaires (F1 pondéré à 0.62). Ces résultats montrent un impact de la qualité du modèle à la fois sur l'aspect naturel des conversations et la fidélité clinique concernant la reconstruction des questionnaires. Une analyse plus approfondie sur un corpus de conversations plus grand permettrait de confirmer cette tendance.

Différentes perspectives sont envisagées dans la continuité de ce travail de recherche, l'objectif final étant d'entraîner un modèle local permettant le remplissage automatique du questionnaire QIDS à partir de conversations réelles entre patient et médecin. Concernant les données générées, une évaluation humaine par des professionnels de santé de l'aspect naturel des dialogues générés est en cours pour valider les données produites. Dans cette même perspective, une évaluation pourrait être réalisée sur les dialogues générés pour mesurer l'influence effective des traits de personnalité sur le style linguistique (par exemple, verbosité, expression émotionnelle),les schémas d'interaction (par exemple, ouverture, coopération) et la manière de formuler les symptômes. Le but est de réussir à produire des conversations plus naturelles qui couvrent l'ensemble des diagnostics, jusqu'aux plus sévères, et pas uniquement les pathologies modérées.

Références

CORACI D., MACCARONE M., REGAZZO G., ACCORDI G., PAPATHANASIOU J. & MASIERO S. (2023). ChatGPT in the development of medical questionnaires. The example of the low back pain. *European Journal of Translational Myology*, **33**.

GAO C., LAN X., LU Z., MAO J., PIAO J., WANG H., JIN D. & LI Y. (2023). S3 : Social-network simulation system with large language model-empowered agents. *arXiv preprint arXiv :2307.14984*.

HOUSSOU R., AUGUSTIN M.-C., RAPPOS E., BONVIN V. & ROBERT-NICOUD S. (2022). Generation and simulation of synthetic datasets with copulas. *arXiv preprint arXiv :2203.17250*.

JIANG H., ZHANG X., CAO X., BREAZEAL C., ROY D. & KABBARA J. (2024). Personallm : Investigating the ability of large language models to express personality traits. In *Findings of the association for computational linguistics : NAACL 2024*, p. 3605–3627.

LEITNER K., CUTRI-FRENCH C., MANDEL A., CHRIST L., KOELPER N., MCCABE M., SELTZER E., SCALISE L., COLBERT J. A., DOKRAS A., ROSIN R. M. & LEVINE L. D. (2025). A Conversational Agent Using Natural Language Processing for Postpartum Care for New Mothers : Development and Engagement Analysis. *JMIR AI*, **4**.

LITTLE R. J. & RUBIN D. B. (2002). Statistical analysis with missing data. john wiley & sons. *New York*.

MCCRAE R. R. & JOHN O. P. (1992). An introduction to the five-factor model and its applications. *Journal of personality*, **60**(2), 175–215.

PAS M. E. T., RUTTEN M. M., BOUWMAN P. R. A., BUISE P. M. P. & DEPARTMENT A. (2020). User Experience of a Chatbot Questionnaire Versus a Regular Computer Questionnaire : Prospective Comparative Study. *JMIR Medical Informatics*, **8**.

RUSH A. J., TRIVEDI M. H., IBRAHIM H. M., CARMODY T. J., ARNOW B., KLEIN D. N., MARKOWITZ J. C., NINAN P. T., KORNSTEIN S., MANBER R. *et al.* (2003). The 16-item quick inventory of depressive symptomatology (qids), clinician rating (qids-c), and self-report (qids-sr) : a psychometric evaluation in patients with chronic major depression. *Biological psychiatry*, **54**(5), 573–583.

TAYAL A., SALUNKE D., DI EUGENIO B., ALLEN-MEARES P., ABRIL E. P., GARCIA-BEDOYA O., DICKENS C. & BOYD A. (2025). Towards conversational assistants for health applications : using chatgpt to generate conversations about heart failure. In *Proceedings of the 26th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, p. 527–537.

TOUDESHKI F. G., JOLIVET P., DURAND-SALMON A. & LIEDNIKOVA A. A. (2021). Zero-Shot Clinical Questionnaire Filling From Human-Machine Interactions. *Proceedings of the 3rd Workshop on Machine Reading for Question Answering*.