Deployment Risk Probe Framework (DRPF): Complementing CER with Actionable Failure Modes
部署風險探針框架（DRPF）：以可處置錯誤型態補充 CER

Ming-Kun Chou[1,2], **Yu-Tai Lo**[3,4], **Wen-Hsiang Lu**[2]

[1]Cross College Elite Program, National Cheng Kung University
[2]Department of Computer Science and Information Engineering, National Cheng Kung University
[3]Department of Geriatrics, National Cheng Kung University Hospital
[4]Department of Public Health, National Cheng Kung University

AN4096750@gs.ncku.edu.tw, n105218@mail.hosp.ncku.edu.tw, whlu@mail.ncku.edu.tw

摘要

高風險場景的語音辨識（ASR）若只比字元錯誤率（CER），往往無法區分後續該排除無語音誤輸出、要求使用者重說，或人工核對辨識內容。本文提出一種分層評估方法：將錯誤分為無語音卻輸出文字、有語音卻無輸出、有語音但內容錯誤，並分別對應上述三種後續處理。我們以預立醫療照護計畫（ACP）對話為案例，比較雲端與邊緣兩套部署流程：當一般診間句 CER 接近時，短答錯誤卻分別以內容錯誤與無輸出為主；其餘條件固定時，僅改語言設定，否定短答的醫療處置詞改寫與詞彙覆蓋亦隨之改變。因此，CER 接近並不表示後續處理方式相同。

Abstract

When deploying ASR in high-stakes settings, character error rate (CER) alone often cannot indicate whether to exclude spurious output on non-speech input, request a repeat utterance, or manually verify recognition content. We propose a layered evaluation that splits errors into text on non-speech input, empty output on speech, and incorrect content on speech, each mapped to one of these follow-up responses. We evaluate on advance care planning dialogue by comparing a cloud and an edge deployment pipeline: when CER is similar on clinic-style sentences, short answers diverge toward content errors versus empty outputs; with other settings held fixed, changing only the language setting also changes medical-term rewrites of negated short answers and lexical coverage. Thus similar CER need not imply the same follow-up handling.

關鍵字：語音辨識、部署評估、預立醫療照護計畫、字元錯誤率、高風險對話

Keywords: Automatic Speech Recognition, Deployment Evaluation, Advance Care Planning, Character Error Rate, High-Stakes Dialogue

1 緒論

部署高風險場景的語音辨識（ASR）時，決策常同時牽涉雲端或邊緣方案的選用、語言設定與語音活動偵測（VAD）設置。當錯誤發生時，若只比較一般句的字元錯誤率（CER），容易把「錯誤率接近」解讀成「後續該做的處置也一樣」。本文提出部署風險探針框架（Deployment Risk Probe Framework, DRPF），依輸入分層將錯誤記為 (i) 無語音卻輸出文字、(ii) 有語音卻無輸出、(iii) 有語音但內容錯誤，並分別對應過濾輸入、重新提問與人工核對等措施。CER 仍適合作為一般轉錄品質的情境指標；DRPF 並不取代 CER，而是避免把它單獨當成部署處置的依據。文獻已指出靜音或非語音輸入可能使 ASR 產生無根據文字 (Koenecke et al., 2024; Barański et al., 2025)，但實務評估仍較少分層計量上述錯誤型態，也較少在短答情境單獨報告可查核的高風險改寫。

預立醫療照護計畫（Advance Care Planning, ACP）對話含確認、否定、猶豫／填充語音與專有名詞，照護引導亦需文化適切 (Schickedanz et al., 2009; Lo et al., 2025; Scheerens et al., 2021)，因此適合作為高風險案例。本文以 ACP-DRP 實作 DRPF，用以計量、比較各類錯誤與對應處置；重點在評估設計，而非單一 ACP 產品的效能報告。

DRPF 架構見圖 1–2。通用核心為：(1) 輸入分層；(2) 有語音卻無輸出與有語音但內容錯誤的分解；(3) 與工程處置對應的指標。ACP-DRP 在靜音、背景噪音、極短語音與猶豫／

Proceedings of the 38th Conference on Computational Linguistics and Speech Processing (ROCLING 2026), pages 77–85
November 6-7, 2026 ©2026 Association for Computational Linguistics

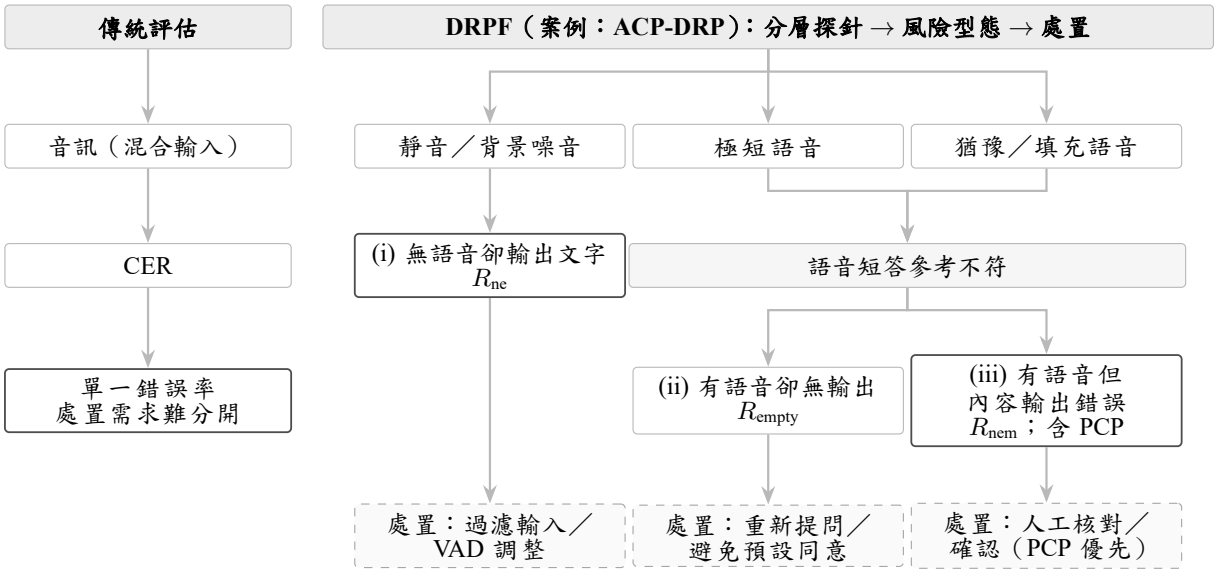

圖 1 左：傳統方法彙整為單一 CER。右：DRPF 依輸入分層給出 (i)/(ii)/(iii)，並對應過濾輸入、重新提問與人工核對，使一般錯誤率與短答處置需求可分開檢視。

填充語音上計算 (i)/(ii)/(iii)，其中 (iii) 含極性衝突探針 PCP（定義見第 3.2 節），並以 CER、關鍵詞漏檢率（KW-ERR）與台語詞彙命中率作情境對照。

為檢驗一般集 CER 相近時短答處置是否仍可能不同，以及流程設定是否改變風險型態，本文提出下列研究問題：

- **RQ1**：在一般診間對照集 C5（醫師提供句，見表 1）的 CER 對照之外，DRPF 是否能在整合流程層級揭露極短語音與猶豫／填充語音中不同的錯誤型態組成，例如以有語音卻無輸出為主或以有語音但內容錯誤為主？
- **RQ2**：在固定其餘設定下，僅變更語言設定或僅變更 VAD，是否分別伴隨不同風險類別分佈的差異？

RQ1 以完整整合流程為描述性比較單位；RQ2 為其餘條件固定、僅變更單一設定時的觀察。本文貢獻如下：

1. 提出 DRPF：以分層錯誤型態補充 CER 的部署導向評估設計。
2. 將錯誤拆解為 (i)/(ii)/(iii)，並分別對應過濾輸入、重新提問與人工核對；以 PCP 示範一類需優先人工核對的短答極性衝突子集。
3. 在 ACP 案例中顯示：一般集 C5 的 CER 可比較時，短答錯誤組成與所需處置仍可能不同；僅變更語言設定或 VAD 時，風險類別分佈亦可能不同。
4. 釐清跨場域固定結構與依場域替換參數（表 3），並以跨語者一致性、Hold-out 與跨域迷你探針檢查替換機制。

2 相關研究

2.1 ASR 產生無根據內容與非典型輸入

端對端 ASR 在靜音或非語音輸入下，可能產生無根據文字 (Koenecke et al., 2024; Barański et al., 2025)；相關文獻亦以 hallucination 指稱語意與輸入無關但仍流暢的輸出，且標準錯誤率不易與其區分 (Frieske and Shi, 2024)。本文量測名稱採「無語音卻輸出文字」，僅在相關研究中對照 hallucination 文獻。既有工作多以「是否產生無根據句子」為判準，較少提供可分別處置的分層指標；短答中的意願扭轉改寫亦少有單獨計數。DRPF 因而分別計算無語音卻輸出文字、有語音卻無輸出與有語音但內容錯誤，並在否定短答另行報告極性衝突探針 PCP，同時以語言設定或 VAD 的對照比較檢驗設定是否改變風險型態。

2.2 部署方式、評估指標與語言覆蓋

Whisper (Radford et al., 2023) 與可串流的 Nemotron (NVIDIA, 2026)（本文以批次模式執行）分別代表常見的雲端與邊緣流程。語言設定（lang）、語音活動偵測（VAD）與簡繁轉換（如 OpenCC）等對行為的影響可能不亞於模型權重，但模型比較中常未完整說明。CER 不易反映臨床等高風險錯誤 (Shor et al., 2023; Whetten and Kennington, 2023)；關鍵詞錯誤亦常為平均錯誤率所掩蓋 (Afonja et al., 2024)。

子集	角色	定義	N
C1	(i)	數位靜音 3/5/8/10 s×25	100
C2A	(i)	醫院底噪；VAD 初篩（內部）	100
C2B	(i)	DEMAND 五場景（可重現）	100
C3	(ii)(iii)/PCP	確認／否定／短答；裁長靜音	200
C3-H	Hold-out	否定短答；與 C3 文字不重複	120
C4	(ii)(iii)	猶豫／填充語音	200
C5	對照	醫師提供診間句＋詞表 $\mathcal{K}$	200
Finance	跨域	金融短答；替換腳本與 $\mathcal{T}_{\text{finance}}$；檢視 Prior-leak§	120‡
Law	跨域	法律短答；替換腳本與 $\mathcal{T}_{\text{law}}$；檢視 Prior-leak§	120‡

表 1 ACP-DRP 各子集說明。四名語者各錄一趟，每人 240 句；其中只有 C5 與詞表 $\mathcal{K}$ 由醫師提供。C2B 取自 DEMAND (Thiemann et al., 2013) 的五個場景（含 PCAFETER 等）。‡Finance、Law 各 N=120；PCP 分母為各域否定短答（N_{neg}=72）。§Prior-leak：金融／法律短答的輸出命中 ACP 觸發詞（見第 3.2 節）。

DRPF 以關鍵詞漏檢率（KW-ERR）與 PCP 補足此限制。

2.3 評估方法論：從字元錯誤到部署決策

CER 與後續任務成效的關聯並不穩定 (Favre et al., 2013)；語意或任務導向指標往往更能反映下游表現 (Rugayan et al., 2023)。高風險情境尚需單獨呈現嚴重錯誤 (Shor et al., 2023; Whetten and Kennington, 2023)；近期研究亦指出 CER 難以分辨 hallucination 重度不同的 ASR 模型 (Frieske and Shi, 2024)。完整對話／代理評估成本較高，且常將 ASR 與特定策略綁定，不利於整合流程迴歸測試。DRPF 定位為以分層錯誤型態補充 CER 的評估設計：新增 (i)/(ii)/(iii) 統計、短答高風險子集探針，以及設定對照比較，與既有 hallucination taxonomy 並用而非取代。框架本身不綁定特定對話系統；本稿以 ACP-DRP 為案例實作。

3 部署風險探針框架（DRPF）

3.1 探針層次與指標

圖 1 將輸入分為三類，對應三類可計數風險類別。第 i 段的 ASR 輸出與參考分別記為 $\hat{y}_i$、y_i；空字串、僅含空白的字串及伺服器特殊標記（例如 `<{silent}>`）一律正規化為空輸出 ε。其餘輸出在相同文字正規化後採完全字串比對，不採語意等價判定。N 為當次計算所用子集大小；$|\{\cdot\}|$ 表示符合條件的片段數。文中比例皆以 $[0,1]$ 定義，表格另以百分比呈現。

案例實作 ACP-DRP 的輸入子集與各指標對應見表 1；後文以 C1–C5 指稱這些子集。C5 為醫師提供之一般診間對照句；$\mathcal{K}$ 為僅用於關鍵詞漏檢（KW-ERR，定義見第 3.3 節）的醫師詞表。Finance/Law 之 Prior-leak 指非 ACP 參考的輸出卻命中 ACP 觸發詞（定義見第 3.2 節）。

無語音卻輸出文字比率僅用於 C1/C2A/C2B；此時所有參考滿足 $y_i = \varepsilon$，且 N 為該非語音子集大小：

$$R_{\text{ne}} = \frac{1}{N}\big|\{i : \hat{y}_i \neq \varepsilon\}\big|. \tag{1}$$

R_{ne} 為無語音輸入上仍輸出文字的比例，對應過濾輸入／語音活動偵測（VAD）調整。例如 5 s 靜音或室內噪音若輸出「你好」等非空文字，即計入一例；輸出為空則不計（表 2）。

有語音卻無輸出比率用於 C3、C4 或兩者聯集；此時本節語音子集皆滿足 $y_i \neq \varepsilon$，且 N 為所選語音子集大小：

$$R_{\text{empty}} = \frac{1}{N}\big|\{i : \hat{y}_i = \varepsilon \wedge y_i \neq \varepsilon\}\big|. \tag{2}$$

R_{empty} 為有語音卻無輸出的比例，對應重新提問與避免預設同意。例如參考為「嗯」卻得到空輸出或 `<{silent}>`，即計入一例。

有語音但內容錯誤比率同樣用於上述語音子集，且僅在參考非空時計算：

$$R_{\text{nem}} = \frac{1}{N}\big|\{i : y_i \neq \varepsilon \wedge \hat{y}_i \neq \varepsilon \wedge \hat{y}_i \neq y_i\}\big|. \tag{3}$$

定義中之 $y_i \neq \varepsilon$ 明確排除非語音輸入，使 R_{nem} 與 R_{ne} 不會在同一片段上重複計數。R_{nem} 為有語音但內容錯誤的比例，對應人工核對與確認機制。它計量需進入人工核對的片段比例，而非語意錯誤率或傷害分數：一般同音字替換與「不要了 → 服藥了」類意願扭轉，在計數上同屬一例；「嗯啊 → 二十」「不要 → 腦要」亦同。表 2 彙整 (i)/(ii)/(iii) 與後文 PCP 的邊界例。

在 $y_i \neq \varepsilon$ 的前提下，R_{empty} 與 R_{nem} 互斥且窮盡所有 $\hat{y}_i \neq y_i$，故 C3/C4 的句級**不符率**為

$$R_{\text{mismatch}} = R_{\text{empty}} + R_{\text{nem}} = \frac{1}{N}\big|\{i : \hat{y}_i \neq y_i\}\big|. \tag{4}$$

R_{mismatch} 為句級完全比對下與參考不同的片段比例，且等於 $R_{\text{empty}} + R_{\text{nem}}$；CER 則為語料字

輸入／參考	輸出	判定
5 s 靜音	空／ `<{silent}>`	(i) 無語音卻輸出文字：否
室內背景噪音	空	(i) 無語音卻輸出文字：否
「不要了」	「服藥了」	(iii)；**PCP**
「好」	「好」	相符
「嗯」	空	(ii) 有語音卻無輸出
「嗯啊」	「二十」	(iii) 非 PCP
「不要」	「腦要」	(iii)；非 PCP

表 2　計分／PCP 邊界例（PCP 操作定義見第 3.2 節）。

元編輯距離（含插入、刪除與替換）除以參考字元總數，兩者不可互換。PCP 自有語音但內容錯誤的案例中再篩高風險子集，不與無輸出重計。相較於以 hallucination 泛稱所有「無根據內容」，本文以操作型之「無語音卻輸出文字」計量 (i)，並將短答極性衝突另以 PCP 呈現，以免混淆不同處置需求。

3.2　極性衝突探針 PCP：示範一類高風險短答反轉

PCP（polarity-conflict probe）不取代完整語意錯誤偵測，而是示範一類需優先人工核對的短答反轉：否定短答被改寫為含高風險處置詞，且輸出仍流暢、易被下游模組當成有效意圖。一旦觸發，確認畫面可能呈現與使用者意願方向相反的內容；在假設部署流程中，這類輸出可優先進入人工核對。PCP 自 R_{nem} 中以可機器執行的字串規則篩出高風險子集；ACP 案例須同時符合：(1) 樣本屬預標記否定／拒絕集 S_{neg}（C3 共 19 句）；(2) 正規化後輸出非空且不等於參考；(3) 輸出命中觸發詞集 $\mathcal{T}_{\text{ACP}} = \{\text{服藥}, \text{插管}, \text{急救}\}$ 至少一項。另報僅計「服藥」的 **PCP-narrow**。例如「不要了」→「服藥了」列入 PCP；「不要」→「腦要」未命中 $\mathcal{T}$，不列入。比率為

$$R_{\text{pcp}} = n_{\text{PCP}} / N_{\text{neg}}. \tag{5}$$

其中 N_{neg}=76 為 19 個否定腳本由四名語者各念一次所得。雲端流程之語言設定參數 `lang` 以 `lang-med` 為主設定、`lang-base` 為對照（流程細節見「案例整合流程」）；雲端 +`lang-med` 為 20/76，約 26.3%（其中 PCP-narrow 19/76），`lang-base` 與邊緣流程皆為 0/76；邊界例見表 2。

窄探針便於部署與迴歸：$\mathcal{T}_{\text{ACP}}$ 為預先選定的 ACP 高風險處置動作詞最小封閉集，與關鍵詞漏檢所用之醫師詞表 $\mathcal{K}$（見第 3.3 節）分開維護。盤點發展集 S_{neg} 顯示觸發以「服藥」為主（19/20），「插管／急救」偶發共現，且後二者亦見於否定腳本；確認可機器執行後即凍結，其後在文字不重複的 Hold-out 上計分且禁止再改 $\mathcal{T}$，亦不依測試集逐條加詞。未命中 $\mathcal{T}$ 的語意扭轉（如「不要抗生素 → 要抗生素」）有意不計；PCP 偏向精確率導向的高風險子集，刻意低估真實語意反轉，以便先抓住必須進人工核對的那一類。規則僅檢查輸出是否命中 $\mathcal{T}$，不要求觸發詞為新增；故若參考本身已含 $\mathcal{T}$ 詞，內容錯誤亦可觸發。ACP 主結果受此影響較小，因觸發以「服藥」為主，而「服藥」不在否定腳本中；PCP-narrow 即對應此較乾淨的子集。

Hold-out 與跨領域。 ACP Hold-out 為與 C3 文字不重複的 30 句否定短答／語者，四人合計 N=120；雲端 +`lang-med` 在 Hold-out 上的 ACP 域 PCP 為 24/120，即 20.0%，與發展集 20/76，約 26.3% 方向一致，降低「規則僅對發展集量身打造」的疑慮。另以金融／法律迷你短答換腳本與觸發集：金融 $\mathcal{T}_{\text{finance}}$ 含轉帳／授權／付款等交易處置詞，法律 $\mathcal{T}_{\text{law}}$ 含簽署／蓋章／契約等法律效力相關詞。三域 PCP 邏輯一致：否定輸入、有語音但內容錯誤、且命中該域預先指定的高風險觸發詞。域內 PCP 分母為各域否定短答（18 句 ×4 語者，N_{neg}=72）；**Prior-leak**（非 ACP 參考卻命中 $\mathcal{T}_{\text{ACP}}$）則以該域全部短答計（各 N=120）。雲端 +`lang-med`：Finance PCP 10/72、Prior-leak 14/120；Law PCP 6/72、Prior-leak 12/120（表 7）。金融與法律的否定集、觸發集與計分規則亦在計分前固定；其目的僅為檢查替換機制能否執行，本稿未另做跨域 development/hold-out 驗證，亦未主張依跨域輸出錯例再調整觸發集。

3.3　情境對照：CER 與 KW-ERR

C5 作為一般診間對話的對照：句與關鍵詞由臨床醫師提供以提升表面效度；其餘探針為研究設計腳本。情境對照另報國語 CER、KW-ERR 與台語詞彙命中率。KW-ERR 採固定詞表 $\mathcal{K}$，$\text{occ}(t, \mathcal{K})$ 為正規化後子字串命中的詞集合（各詞至多一次）；式 (6) 只計參考應出現卻未出現在輸出的詞：

$$\text{KW-ERR} = \frac{\sum_i |\text{occ}(y_i, \mathcal{K}) \setminus \text{occ}(\hat{y}_i, \mathcal{K})|}{\sum_i |\text{occ}(y_i, \mathcal{K})|}. \tag{6}$$

故 KW-ERR 為偏保守的關鍵詞漏檢率。固定詞表 $\mathcal{K}$ 由臨床醫師提供的診間 ACP 常用詞整理而成，共 49 詞，含「預立醫療決定」「鼻胃管」「葉克膜」等。此詞表**僅**用於 KW-ERR，不併入 PCP 的 $\mathcal{T}_{\text{ACP}}$。台語詞彙命中率以朗讀子集中逗號分隔的詞彙標註為參考：經與 CER 相同之文字正規化後，以子字串比對計算命中

詞數／參考詞總數（keyword recall）；因參考非逐字稿，台語覆蓋不以 CER 單獨解讀。C5／KW-ERR 使用同一組四語者、醫師提供句共 200 句（表 6、8、10）。報告時同時呈現錯誤型態探針與 CER／關鍵詞對照。

3.4 可泛化性：固定核心與依場域探針

DRPF 不綁定 ACP；跨場域可固定的是通用核：輸入分層（非語音／極短／猶豫或低確信語音）、(i)/(ii)/(iii) 分解（R_{ne}、R_{empty}、R_{nem}），以及與過濾輸入／重新提問／人工核對的處置對應；CER 與關鍵詞漏檢作為情境對照的角色亦同。依場域替換的是腳本內容、否定集 S_{neg}、觸發詞集 $\mathcal{T}$ 與關鍵詞表 $\mathcal{K}$；這些需重寫，但不改框架結構。

映射示例：(a) 法律語音代理：否定短答配 $\mathcal{T}_{\text{law}}$（如簽署／蓋章／契約），監測「拒絕短答後出現簽署、蓋章或契約等法律效力相關詞」；(b) 金融客服：否定／撤銷短答配 $\mathcal{T}_{\text{finance}}$（如轉帳／授權／付款），監測「拒絕被改寫為授權轉帳等交易處置詞」；(c) 教育口試或客服 IVR：以確認／否定短答與專有名詞表替換 C3 ／ $\mathcal{K}$，C1–C2 非語音層可直接重用。表 3 將上述邊界整理為可檢查的操作規則：何者跨場域固定、何者依部署場域替換。

本稿的金融／法律迷你探針旨在驗證計分管線的可替換性，並檢視各域 PCP 觸發與 ACP Prior-leak，而非估計跨域錯誤率或外推絕對風險。因此，可泛化性主張限於框架結構、計分程序與探針替換機制。

3.5 ACP-DRP：DRPF 的案例實作

假設部署流程用來錨定處置對應：使用者回答 ACP 引導問題，系統先轉寫再呈給使用者或照護專業者確認；空輸出觸發重新提問，有語音但內容錯誤進入人工核對，無語音卻輸出文字應在輸入端阻擋。否定短答若被改寫為含「服藥」等指示詞，確認畫面可能與意願相反，故以 PCP 單獨監測。

如表 1 所示，ACP-DRP 依層次設定對應指標：C2B 可公開重現、C2A 僅供內部；C3 裁長靜音、C4 保留猶豫／填充語音本身。C5 句與 $\mathcal{K}$ 由臨床醫師提供；C3 ／ Hold-out ／ C4 ／金融／法律為研究設計腳本。PCP 的 $\mathcal{T}$ 不與 $\mathcal{K}$ 混用。後續實驗以這些子集檢驗 RQ1 與 RQ2。

3.6 案例整合流程

雲端流程：使用遠端 Whisper large-v3-turbo (OpenAI, 2024; Radford et al., 2023)。此 API 提供語言設定參數 `lang`；主設定 `lang-med` 與對照 `lang-base` 已於第 3.2 節報

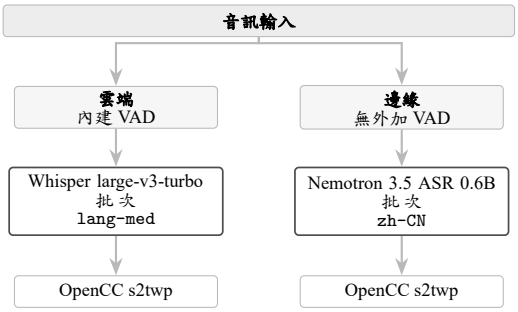

圖 2 案例整合流程，比較單位為整合流程而非單一模型；設定比較見第 5.2 節。

告 PCP 結果，兩者在台語詞彙命中率、國語 CER、C5 CER、KW-ERR 與短答探針結果上呈現不同取捨；因 ACP 對話可能夾雜台語，本文以台語詞彙覆蓋較佳的 `lang-med` 為主設定。第三方服務的內部語言覆蓋與解碼機制未公開。雲端流程使用內建 VAD，辨識結果再經 OpenCC (Kuo and OpenCC contributors, 2010) s2twp 轉換。**邊緣流程**：使用 Nemotron 3.5 ASR 0.6B (NVIDIA, 2026) 與 `zh-CN` 語言設定，辨識結果同樣經 OpenCC 轉換；主要設定不另加 VAD。此外另使用公開 Whisper 與門檻為 0.5 的 Silero VAD 做對照比較。

圖 2 標出可能影響結果的因素，所以結果應解讀為完整整合流程的表現。對照比較包括：雲端流程僅變更語言設定參數 `lang`，以及公開 Whisper 與邊緣流程分別加入或移除 VAD。

4 實驗設定

4.1 語料與文字格式統一

國語語料取自實驗室 TTS 合成之繁體中文子集，原始取樣率為 22 kHz，重新取樣至 16 kHz；以亂數種子 42 抽取 500 句，作為辨識表現對照。台語語料則自同一語料庫之朗讀子集抽取 500 句，並以台語詞彙命中率作為詞彙覆蓋的概略指標：在 `lang-med` 設定下，台語詞彙命中率為 86.1%、CER 為 9.0%；國語／台語 500 句為既有對照，非本次人聲重錄。人聲探針為四名成人（s1、s3 男；s2、s4 女），在安靜室內依腳本錄製 C3、Hold-out、金融、法律、C4、C5，每人 240 句；語者均知情同意錄音僅供研究評估使用，本研究未涉及臨床介入或病患個資。C5 句與關鍵詞由臨床醫師提供；其餘短答／猶豫探針為研究設計腳本。醫院場域底噪 C2A 以環境音為目標採集，不以可辨識人聲或診療對話為收錄對象。CER 計算前統一全形與半形字元、移除標點與空白，並轉為繁體中文；各整合流程均採相同文字正規化程序。

層次	跨場域固定	依場域替換
輸入層	非語音、極短語音、低確信／猶豫語音的分層	場域可接受的音訊條件與腳本
計分層	R_{ne}、R_{empty}、R_{nem} 及 (ii)/(iii) 分解	文字正規化與空輸出特殊標記
風險探針層	有語音但內容錯誤中另設高風險子集的原則	否定集 S_{neg}、觸發集 $\mathcal{T}$、關鍵詞表 $\mathcal{K}$
處置層	過濾輸入、重新提問、人工核對的對應邏輯	實際工作流程、覆核門檻與升級規則

表 3 DRPF 的可移植邊界：核心分層、計分與處置對應固定；腳本與領域詞表依部署場域替換。

4.2 ACP-DRP 建置摘要

C1 為依長度分組的純數位靜音。C2B 自 DEMAND 的 PCAFETER、OHALLWAY、OOFFICE、PRESTO、DLIVING 五場景各取 20 段。C2A 為醫院背景噪音，經 VAD 初篩且僅供內部使用；公開重現與對外主張的 (i) 以 C1/C2B 為準。C3/C4 ／ Hold-out ／ Finance ／ Law ／ C5 由四名語者依相同腳本各錄一趟；音檔皆轉為 16 kHz 單聲道。C3 另裁除前後較長的靜音。C3 腳本共 50 句，依類別劃分為否定／拒絕 19 句、確認含詢問 24 句、其他短答 7 句。C3 與 C4 合計 400 段，其中 19 句否定腳本由四名語者各念一次得 76 段，用於計算 PCP；Hold-out ／ Finance ／ Law 另計（表 1）。設計時優先確認探針能否區分不同風險類別，以及流程設定變更後是否呈現相應差異。

4.3 整合流程與設定比較

主要比較為雲端流程 +lang-med 與邊緣流程。其餘條件固定下的對照：(a) 雲端僅改 lang（lang-med↔lang-base）；(b) 公開 Whisper±VAD；(c) 邊緣 ±VAD。公開 Whisper±VAD 用來觀察 VAD 與各風險類別及 C5 CER 的共變。設定摘要見表 4。所有比較在同一批音檔與同一計分腳本上執行；主表取單次輸出以模擬部署時可觀察行為。RQ2 為本流程之對照觀察，不宜推論為一般因果效應。

項目	雲端主設定	邊緣主設定
模型	Whisper large-v3-turbo，遠端 API	Nemotron 3.5 ASR 0.6B
語言	lang-med；對照 lang-base	zh-CN
VAD	API 內建	無外加；對照 +Silero 0.5
後處理	OpenCC s2twp	OpenCC s2twp
比較單位	整合流程；主表單次固定結果	同左

表 4 設定摘要。公開 Whisper±VAD 見表 9；雲端 API 內部解碼／版本細節不可查核。

5 實驗結果與討論

公開 Whisper 加上 VAD 後，無語音卻輸出文字由各子集 100/100 降為 0/100，短答不符組成與 C5 CER 亦一併變動（表 9、10）。同音檔僅改雲端 lang 時，PCP 由 20/76 變 0/76，台語詞彙命中率由 86.1% 降至 11.2%。醫師提供的 C5 上，兩主流程 CER 為 11.2% 與 7.6%，短答則分別出現有語音但內容錯誤 243/400 與有語音卻無輸出 72/400。分項結果如下。

5.1 RQ1：DRPF 區分一般集 CER 下的短答風險型態

元件層級的對照觀察見 RQ2。表 5 顯示兩套流程在 C1/C2B 及內部 C2A 上均未產生文字，但在極短語音與猶豫／填充語音上的風險型態明顯不同。

設定	(i) 各子集	不符	(ii)	(iii)
雲端 +lang-med	0/100	259/400	16/400	**243/400**
雲端 +lang-base	0/100	128/400	16/400	**112/400**
邊緣	0/100	132/400	**72/400**	60/400

表 5 RQ1 探針結果。(i) 無語音卻輸出文字：C1/C2A/C2B 各 N=100 仍為既有結果 0/100；(ii) 有語音卻無輸出、(iii) 有語音但內容錯誤，皆為 C3+C4、N=400（四語者）。粗體標示各列在 (ii)/(iii) 中較大者，亦即該列主導的短答錯誤型態。

lang-med 的不符率為 64.8%，即 259/400，其中以有語音但內容錯誤為主（243/400）；邊緣不符率為 33.0%，即 132/400，其中以有語音卻無輸出為主（72/400）。lang-base 在相同音檔上不符率 128/400、PCP 0/76。對照表 6，醫師提供 C5 上 CER 雖為 11.2% 與 7.6%，短答所需處置卻不相同：前者偏人工核對，後者偏重新提問。本樣本 McNemar 僅摘要同一錄音上之句級配對方向（連續性校正 χ^2；非獨立語者推論，見第 5.4 節）：C3+C4 有語音但內容錯誤上 lang-med 對邊緣 b=198、c=15（χ^2=155.51），對 lang-base b=141、c=10（χ^2=111.92）；PCP 20/76 對 0/76（χ^2=18.05）。不一致格遠大於反向，方向與表 5 一致；因語者內重複觀測，此處不將檢定結果解讀為跨族群顯著性。整體不符集中於 C4：雲端 C4 有語音但內容錯誤 128/200，邊緣 C4 有語音卻無輸出 57/200；PCP 集中於 C3 ／ Hold-out。

跨語者一致性與 Hold-out 檢查。 在四位語者中，lang-med 的 PCP 計數分別為 8/19、1/19、6/19、5/19；相對地，邊緣與 lang-base 在各語者的對應否定短答中均未

觀察到 PCP。文字不重複的 Hold-out 上同規則得 ACP 域 PCP 24/120（20.0%），與發展集 26.3% 方向一致。跨域迷你探針（表 7）在換腳本與換 $\mathcal{T}$ 後仍能跑通同一計分管線。因此，這些檢查支持所觀察的方向並非只由單一語者或單一發展腳本驅動。

整合流程	國語 CER (%)	台語詞彙命中 (%)†	C5 CER (%)	KW-ERR (%)
雲端 +lang-med	6.5	**86.1**	11.2	48.2
雲端 +lang-base	**2.6**	11.2	**3.9**	**14.3**
邊緣	5.6	—	7.6	42.9

表 6 辨識表現對照。國語／台語 500 句為既有對照；C5 ／ KW-ERR 為醫師提供句、四語者合計 N=200。† 僅雲端提供台語設定。粗體為各欄較佳值（CER ／ KW-ERR 較低、詞彙命中較高）。

設定	Finance PCP	Finance leak	Law PCP	Law leak
雲端 +lang-med	10/72	14/120	6/72	12/120
雲端 +lang-base	0/72	0/120	0/72	0/120
邊緣	1/72	0/120	0/72	0/120

表 7 跨域迷你探針。Finance ／ Law PCP 分母為各域否定短答 N_{neg}=72（18×4）；Prior-leak 分母為各域全部短答 N=120。Finance ／ Law PCP 依 $\mathcal{T}_{\text{finance}}$、$\mathcal{T}_{\text{law}}$ 計數；Prior-leak 指非 ACP 參考的輸出命中 $\mathcal{T}_{\text{ACP}}$。數字為事件數而非跨域錯誤率估計；用途是檢查觸發集替換與場域隔離。

有語音卻無輸出較易察覺，因空字串可觸發重新提問；有語音但內容錯誤（含 PCP）則不易察覺，後續模組可能將流暢錯字視為有效意圖。本文未評估完整對話系統，但此一差異說明須分別計算 (ii) 與 (iii)，才能對應不同工程處置。

5.2 RQ2：DRPF 比較整合設定的影響

以下檢驗設定變更是否伴隨風險類別差異：若風險型態隨 lang 或 VAD 改變，部署評估便不能只比模型，尚須把流程設定納入迴歸測試。

就 lang 而言，表 8 顯示僅變更雲端此參數時，台語詞彙命中率、C5 ／ KW-ERR 與短答風險均隨之呈現差異：PCP 20/76 → 0/76，C3+C4 不符 259/400 → 128/400，C5 CER ／ KW-ERR 11.2% ／ 48.2%→3.9% ／ 14.3%。同一 API 下，語言設定伴隨短答風險組成的不同取捨。

lang	台語 CER	詞彙命中	國語 CER	C5 CER	KW-ERR
lang-med	**9.0**	**86.1**	6.5	11.2	48.2
lang-base	88.4	11.2	**2.6**	**3.9**	**14.3**

表 8 僅變更雲端 lang。台語／國語各 500 句為既有對照；台語參考非逐字稿，台語 CER 僅供參考；C5 ／ KW-ERR 為醫師提供句、四語者 N=200。粗體為各欄較佳值。

lang-med 台語詞彙命中率較高（86.1%）且出現 PCP；改 lang-base 後 PCP 消失但命中率降至 11.2%。一個合理假設是：為提高專業詞彙召回，語言先驗可能偏向醫療相關詞，使極短輸入時「不要了／服藥了」易混淆；跨域 Prior-leak 方向一致，惟非解碼器機制證明。部署規格若只訂語言覆蓋下限、不訂 PCP ／ KW-ERR 上限，可能漏掉此取捨。

就 VAD 而言，在公開 Whisper 的非語音子集中，無 VAD 時 C1/C2A/C2B 各為 100/100，加 VAD 後各為 0/100（表 9）。在同一對照比較中，短答不符組成亦有變動：不符由 133/400 增至 145/400，其中有語音卻無輸出由 0/400 增至 16/400、有語音但內容錯誤由 133/400 降至 129/400。此外，公開 Whisper 的 C5 CER 亦由 12.3% 降至 4.4%（表 10）。相對地，邊緣流程外加同一門檻的 Silero VAD 後，非語音子集仍為 0/100，但 C3+C4 不符由 132/400 升至 145/400（有語音卻無輸出 72 → 80、有語音但內容錯誤 60 → 65），C5 CER 亦由 7.6% 微升至 7.8%。因此，VAD 與非語音輸入高度相關，但對短答／一般集的伴隨效應依流程而異；分層指標可把這些變化拆開看，而不必歸成單一數字。雲端流程雖已靠內建 VAD 避免無語音卻輸出文字，短答仍可能出現有語音但內容錯誤與 PCP。

設定	(i) 各子集	不符	(ii)	(iii)
雲端 +lang-base‡	0/100	128/400	16/400	112/400
公開 Whisper§	100/100	133/400	0/400	133/400
公開 Whisper+VAD§§	0/100	145/400	16/400	129/400
邊緣 +VAD§§	0/100	145/400	80/400	65/400

表 9 對照探針結果。‡ 列為四語者主對照；§ ／ §§ 為同一批四語者短答＋既有 C1/C2A/C2B 非語音子集。

設定	CER	KW-ERR
雲端 +lang-med	11.2	48.2
雲端 +lang-base	3.9	14.3
邊緣	7.6	42.9
邊緣 +VAD	7.8	42.0
公開 Whisper	12.3	27.7
公開 Whisper+VAD	4.4	26.8

表 10 C5 的 CER 與 KW-ERR（%）。各列皆為同一組醫師提供句、四語者 N=200。

5.3 錯誤分析與處置對應

除表 2 之邊界例外，表 11 另列 ACP 一般句與設定對照之代表性錯例。

例如「不要了」被改寫為「服藥了」、「不要插管」被改寫為「服藥插管」；此類 PCP 錯誤在 76 筆否定短答中出現 20 筆，語者間為 8/19、1/19、6/19、5/19，且呈現一致的錯誤型態：否定語句被改寫為含 $\mathcal{T}_{\text{ACP}}$ 觸發詞的輸出。對同一錄音檔僅改用 lang-base 重跑後，即可正確辨識為「不要了」；因輸入音檔完全

類型	參考	輸出
ACP	預立醫療照護諮商的流程	玉利醫療招呼知傷者流程
ACP	氣切跟插管哪一個比較痛苦？	剃尺跟柴管哪一個比較痛苦
PCP†	不要了	服藥了
PCP†	不要插管	服藥插管
對照 lang-base	不要了	不要了
猶豫／填充語音	嗯	<{silent}>

表 11　錯例摘要，涵蓋雲端與邊緣流程。†PCP；R_{pcp}=20/76。對照列為同一錄音改用 lang-base 後的結果。

相同，此差異來自流程設定，而非該次錄音條件。相對地，「不要」→「腦要」、「嗯啊」→「二十」等仍屬 R_{nem}，但未命中 $\mathcal{T}_{\text{ACP}}$，所以不計入 PCP，見表 2；由此可見 PCP 為高風險子集，而非唯一的有語音但內容錯誤指標。

各風險類別直接對應工程處置：R_{ne} 偏高 → 調 VAD／過濾輸入；R_{empty} 偏高 → 重新提問且勿預設同意；$R_{\text{nem}}/R_{\text{pcp}}$ 偏高 → 人工核對；KW-ERR 偏高 → 詞表迴歸。同一探針可在變更設定後重跑，作為跨版本迴歸測試。

(iii) 的半自動處置。 本文將 (iii) 對應人工核對，但分層結果亦可用來縮小人工核對的範圍：命中 PCP 或 R_{nem} 的片段可先交由大型音訊語言模型結合領域詞表 $\mathcal{K}$ ／ $\mathcal{T}$（例如以 RAG 檢索或詞表約束解碼）重新判讀，僅在兩次結果不一致或仍命中 $\mathcal{T}$ 時送交人工確認。此類半自動偵測與改寫屬處置端設計，其效益與誤改風險需另行評估；DRPF 的分層計數可作為評估此類機制前後差異的基準。

5.4　效度威脅與研究限制

本稿為 DRPF 的概念驗證（四人腳本樣本），結果解讀為「可區分風險型態」，而非母體缺陷率或臨床安全定案。跨語者一致性、Hold-out 與跨域迷你探針亦不構成跨族群、跨錄音條件或跨版本的穩定性估計。第 5.1 節 McNemar 為句級配對之不一致方向摘要（如 C3+C4 之 N=400），語者數不足以支持跨族群統計推論。語者間 PCP 絕對計數差異大（8/19–1/19）；真實高齡／病患聲學未納入；C4 腳本化猶豫與自然不流暢可能落差最大；C5 雖由醫師提供句，仍非自發診間對話。國語 500 句為 TTS、台語 500 句為朗讀，僅作情境對照，不可外推至臨床自發語音。PCP 觸發集刻意偏窄；計分僅檢查輸出是否命中 $\mathcal{T}$，金融／法律否定腳本中部分參考本身已含對應觸發詞，故跨域 PCP 可含「保留觸發詞的內容錯誤」，不宜直接解讀為極性反轉率——ACP 主結果以「服藥」為主，受此影響較小。RQ1 為整合流程描述性比較；RQ2 的 lang 為可重跑工程現象，非 API 內部機制證明。C2A 僅內部；(i) 對外以 C1/C2B 為準。主表為單次輸出，本稿未報告同一音檔重跑之變異。可查核項目見文末聲明。

對話條件覆蓋。 本稿之短答／猶豫探針以清楚回合、單一語者、非串流為前提。實務高風險部署另涉及即時串流輸入、語者重疊、話輪打斷與界線不清等條件，本稿未納入；本文結論不宜直接外推至上述情境。跨這些條件的錯誤型態分佈是否仍呈現同方向差異，需以能保留串流／重疊音檔的另批語料再測。

ASR 系統覆蓋。 主比較為遠端 Whisper large-v3-turbo 整合流程與 Nemotron 3.5 ASR 邊緣流程，另加公開 Whisper±VAD 對照。較新之大型音訊語言模型（如 Qwen3-ASR）與專為繁體中文訓練之 ASR（如 Breeze-ASR），以及具較強語意先驗之 LLM-based ASR，均未納入評估；本文所觀察之短答風險型態差異是否於此類模型上仍成立，尚待後續補齊。此外，(iii) 有語音但內容錯誤之處置目前對應人工核對；以音訊語言模型結合領域詞彙（例如 RAG 檢索與詞表約束解碼）作半自動偵測或改寫，屬於處置端的延伸，不在本評估設計範圍內。

6　結論與未來工作

本文提出 DRPF，以 (i)/(ii)/(iii) 補充 CER，並對應過濾輸入、重新提問與人工核對。在 ACP-DRP 案例中，一般集 C5 的 CER 可比較時，短答處置需求卻不同（有語音但內容錯誤 vs. 有語音卻無輸出，見表 5）；僅變更 lang 時，極性衝突與詞彙覆蓋亦隨之改變；公開 Whisper 加 VAD 後，無語音卻輸出文字由各子集 100/100 降為 0/100，且 C5 與短答指標一併變動。可遷移者為評估結構（輸入分層與 (ii)/(iii) 分解）；依場域替換腳本與觸發集後，同一計分程序可重跑為案例內／跨設定檢查，CER 保留為情境對照。DRPF 不取代 CER，而是補足 CER 難以區分的部署處置需求；ACP-DRP 為案例實作。

未來工作：(1) 增語者並複核凍結規則下的跨語者一致性；(2) 納入高齡者自然口語、串流輸入、語者重疊與話輪打斷等更困難的對話條件；(3) 將探針接入 LLM 語音代理以觀測端對端效益，並探索以音訊語言模型結合領域詞彙（RAG ／詞表約束解碼）作為 (iii) 之半自動偵測與改寫；(4) 於較新之大型 ASR（如 Qwen3-ASR、Breeze-ASR）重跑相同探針以檢

驗風險型態；(5) 擴充多語言／多場域探針規格。

可重現性聲明

本文釋出下列材料以供結構性複現：(1) C1 靜音（3/5/8/10 s 各 25 段）與 C2B DEMAND 五場景取樣清單；(2) 文字正規化、R_{ne} ／ R_{empty} ／ R_{nem} ／ PCP ／ KW-ERR 之計分程式，以及公開 Whisper、Silero VAD（門檻 0.5）與 OpenCC `s2twp` 之設定檔；(3) C3 短答與 C4 猶豫／填充之逐句腳本；(4) 主表聚合計數見本文正文與各表。含個資或機構限制之材料不釋出，包括 C2A 醫院底噪、語者個人錄音檔、雲端 API 內部 `lang` 細節，以及 Hold-out ／金融／法律腳本與醫師提供之 C5 腳本及關鍵詞表 $\mathcal{K}$。方法結構之複現以 (2) 為準；主表數值受錄音、API 版本與環境差異影響，讀者以自行錄製之音訊重跑所得結果預期同結構、非同值。上述材料見：

`https://github.com/MKChou/DRPF-rocling2026`

致謝

本研究承蒙國家科學及技術委員會（National Science and Technology Council）計畫補助（編號：NSTC 115-2314-B-006-045），特此致謝。

References

Tejumade Afonja, Tobi Olatunji, Sewade Ogun, Naome A. Etori, Abraham Owodunni, and Moshood Yekini. 2024. Performant ASR models for medical entities in accented speech. In *Proc. Interspeech 2024*, pages 2315–2319.

Mateusz Barański, Jan Jasiński, Julitta Bartolewska, Stanisław Kacprzak, Marcin Witkowski, and Konrad Kowalczyk. 2025. Investigation of Whisper ASR hallucinations induced by non-speech audio. In *ICASSP 2025 – 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Benoit Favre, Kyla Cheung, Siavash Kazemian, Adam Lee, Yang Liu, Cosmin Munteanu, Ani Nenkova, Dennis Ochei, Gerald Penn, Stephen Tratz, Clare Voss, and Frauke Zeller. 2013. Automatic human utility evaluation of ASR systems: Does WER really predict performance? In *Proc. Interspeech 2013*, pages 3463–3467.

Rita Frieske and Bertram E. Shi. 2024. Hallucinations in neural automatic speech recognition: Identifying errors and hallucinatory models. arXiv preprint arXiv:2401.01572.

Allison Koenecke, Anna Seo Gyeong Choi, Katelyn X. Mei, Hilke Schellmann, and Mona Sloane. 2024. Careless whisper: Speech-to-text hallucination harms. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pages 1672–1681.

Carbo Kuo and OpenCC contributors. 2010. OpenCC: Conversion between traditional and simplified chinese. GitHub repository. S2twp configuration; accessed 2026-07-05.

Yu-Tai Lo, Sheng-Yu Fan, Chung-Yi Li, Deng-Chi Yang, Chi-Chang Huang, and Mei-Hua Chen. 2025. Prevalence and factors associated with informal advance care planning discussion and do-not-resuscitate directives in patients at geriatric clinics. *Journal of the Formosan Medical Association*. Online ahead of print.

NVIDIA. 2026. Nemotron 3.5 ASR. Hugging Face model card. Streaming multilingual model, 0.6B parameters; accessed 2026-07-05.

OpenAI. 2024. Whisper large-v3-turbo. Hugging Face model card. Finetuned pruned large-v3; approximately 809M parameters; released September 2024.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, pages 28492–28518.

Janine Rugayan, Giampiero Salvi, and Torbjørn Svendsen. 2023. Perceptual and task-oriented assessment of a semantic metric for ASR evaluation. In *Proc. Interspeech 2023*, pages 2158–2162.

Charlotte Scheerens, Joni Gilissen, Aiesha M. Volow, Jana L. Powell, Clarissa M. Ferguson, David Farrell, Brookelle Li, Corinne Berry, and Rebecca L. Sudore. 2021. Developing eHealth tools for diverse older adults: Lessons learned from the PREPARE for your care program. *Journal of the American Geriatrics Society*, 69(10):2939–2949.

Adam D. Schickedanz, Dean Schillinger, C. Seth Landefeld, Sara J. Knight, Brie A. Williams, and Rebecca L. Sudore. 2009. A clinical framework for improving the advance care planning process: Start with patients' self-identified barriers. *Journal of the American Geriatrics Society*, 57(1):31–39.

Joel Shor, Ruyue Agnes Bi, Subhashini Venugopalan, Steven Ibara, Roman Goldenberg, and Ehud Rivlin. 2023. Clinical BERTScore: An improved measure of automatic speech recognition performance in clinical settings. In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 1–7, Toronto, Canada. Association for Computational Linguistics.

Joachim Thiemann, Nobutaka Ito, and Emmanuel Vincent. 2013. DEMAND: a collection of multi-channel recordings of acoustic noise in diverse environments. Zenodo. CC BY-SA 3.0; also described at the 21st International Congress on Acoustics (ICA 2013).

Ryan Whetten and Casey Kennington. 2023. Evaluating and improving automatic speech recognition using severity. In *Proceedings of the 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 79–91, Toronto, Canada. Association for Computational Linguistics.