УДК 004.934.2
doi:10.31799/1684-8853-2026-1-19-35
EDN: DSRKFE

Обзоры
Surveys

Аналитический обзор применения больших языковых моделей для автоматического распознавания речи

И. С. Кипяткова[a], *канд. техн. наук, доцент, старший научный сотрудник, orcid.org/0000-0002-1264-4458, kipyatkova@iias.spb.su*
М. Д. Долгушин[a], *младший научный сотрудник, orcid.org/0000-0002-4344-2330*
И. А. Кагиров[a], *научный сотрудник, orcid.org/0000-0003-1196-1117*
[a]*Санкт-Петербургский Федеральный исследовательский центр РАН, 14-я линия В. О., 39, Санкт-Петербург, 199178, РФ*

Введение: *одной из тенденций в области обработки естественных языков является использование больших языковых моделей. В системах распознавания речи большие языковые модели начинают заменять традиционные благодаря их способности учитывать более широкий контекст.* ***Цель:*** *выполнить систематизацию и обобщение существующих методов совместного использования систем автоматического распознавания речи и больших языковых моделей.* ***Результаты:*** *выявлены основные тенденции внедрения больших языковых моделей в процесс распознавания речи. Анализ продемонстрировал, что применение больших языковых моделей для переоценки гипотез и коррекции ошибок распознавания стабильно улучшает результаты распознавания, хотя это улучшение не всегда является принципиальным и сопряжено с риском генерации недостоверной информации вследствие возможных галлюцинаций моделей. Установлено, что контекстуализация и контекстное обучение больших языковых моделей могут как значительно улучшать, так и, в некоторых случаях, ухудшать результаты распознавания.* ***Практическая значимость:*** *полученные выводы могут найти практическое применение при создании систем автоматического распознавания речи на различных естественных и малоресурсных языках, а также для речи с переключением кодов.* ***Обсуждение:*** *установлено, что рекуррентные и диффузионные архитектуры больших языковых моделей пока не получили широкого распространения в задачах распознавания речи, однако обладают значительным потенциалом. Отмечена тенденция к использованию декодерных архитектур, что в свою очередь порождает проблемы галлюцинаций и ориентации на письменные нормы при генерации текста.*

Ключевые слова *— большие языковые модели, переоценка гипотез, коррекция ошибок, контекстное обучение, автоматическое распознавание речи.*

Для цитирования: Кипяткова И. С., Долгушин М. Д., Кагиров И. А. Аналитический обзор применения больших языковых моделей для автоматического распознавания речи. *Информационно-управляющие системы*, 2026, № 1, с. 19–35. doi:10.31799/1684-8853-2026-1-19-35, EDN: DSRKFE
For citation: Kipyatkova I. S., Dolgushin M. D., Kagirov I. A. Analytical review of the application of large language models for automatic speech recognition. *Informatsionno-upravliaiushchie sistemy* [Information and Control Systems], 2026, no. 1, pp. 19–35 (In Russian). doi:10.31799/1684-8853-2026-1-19-35, EDN: DSRKFE

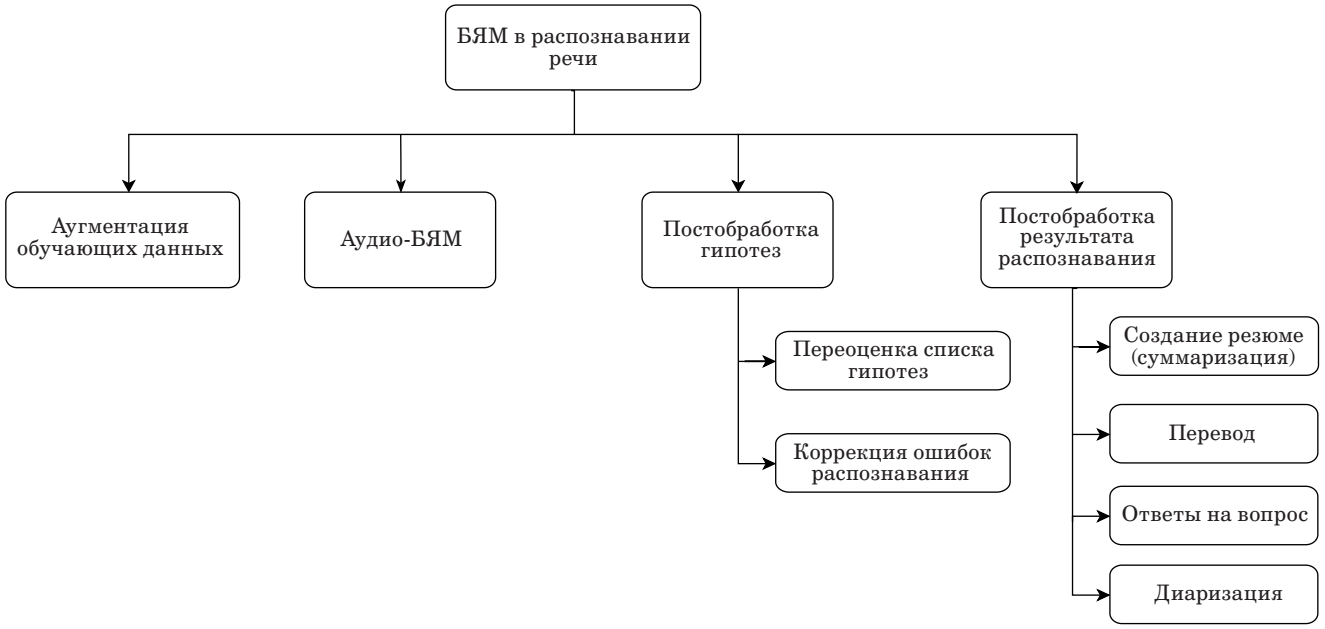
Введение

Большие языковые модели (БЯМ; large language models, LLM) в настоящее время все чаще применяются для решения широкого круга задач из области обработки естественных языков: от обнаружения ошибок в документах до ответов на вопросы по содержанию текста и машинного перевода. При этом БЯМ успешно работают не только с текстовой модальностью, но еще и с аудио- и видеоданными.

Целью настоящего обзора является систематизация и обобщение существующих методов применения БЯМ для распознавания речи. Стратегии применения БЯМ достаточно разнообразны: БЯМ могут использоваться как отдельный модуль для исправления ошибок, допущенных основной системой распознавания речи (СРР), как модуль для переоценки и выбора лучшей гипотезы распознавания или же быть интегрированы в единую архитектуру вместе с речевым кодером (рис. 1). Помимо этого, способности БЯМ к генерации текста оказываются полезными при аугментации обучающих данных. Наконец, БЯМ могут использоваться для различных задач, относящихся к постобработке результатов распознавания, например для генерации ответов в диалоговых системах или для машинного перевода.

Основные архитектуры БЯМ

Большая языковая модель — предобученная языковая модель, которая состоит из нейронной сети со множеством параметров (10 млрд и более), обученной на большом количестве неразмеченного текста или иных данных [1, 2]. В некоторых работах [3] предобученные языковые модели, имеющие менее 10 млрд параметров, но



■ ***Рис. 1.*** Основные области применения БЯМ
■ ***Fig. 1.*** Main application areas of LLMs

демонстрирующие показатели, сопоставимые с таковыми у БЯМ, называются малыми языковыми моделями (small language models).

В большинстве фундаментальных языковых моделей (foundation models) применяется архитектура трансформер [4, 5]. Базовая архитектура трансформер подразумевает два основных блока: кодер, преобразующий входную последовательность в скрытое представление, и декодер, порождающий выходную последовательность с использованием как скрытого представления, так и предыдущих элементов выходной последовательности. Архитектуры современных БЯМ можно разбить на три класса: архитектуры типа «кодер-декодер», «только кодер» и «только декодер». С 2024 г. появились БЯМ, в которых применяются альтернативные архитектуры, например модели непрерывного пространства состояний [6] и диффузионные модели [7].

Архитектура «кодер-декодер» (классическая архитектура модели трансформер) используется для так называемых задач преобразования последовательности в последовательность (sequence-to-sequence), актуальных, в том числе, в рамках машинного перевода и распознавания речи. Подобную архитектуру имеют предобученные модели BART [8], mBART [9], T5 [10] и mT5 [11].

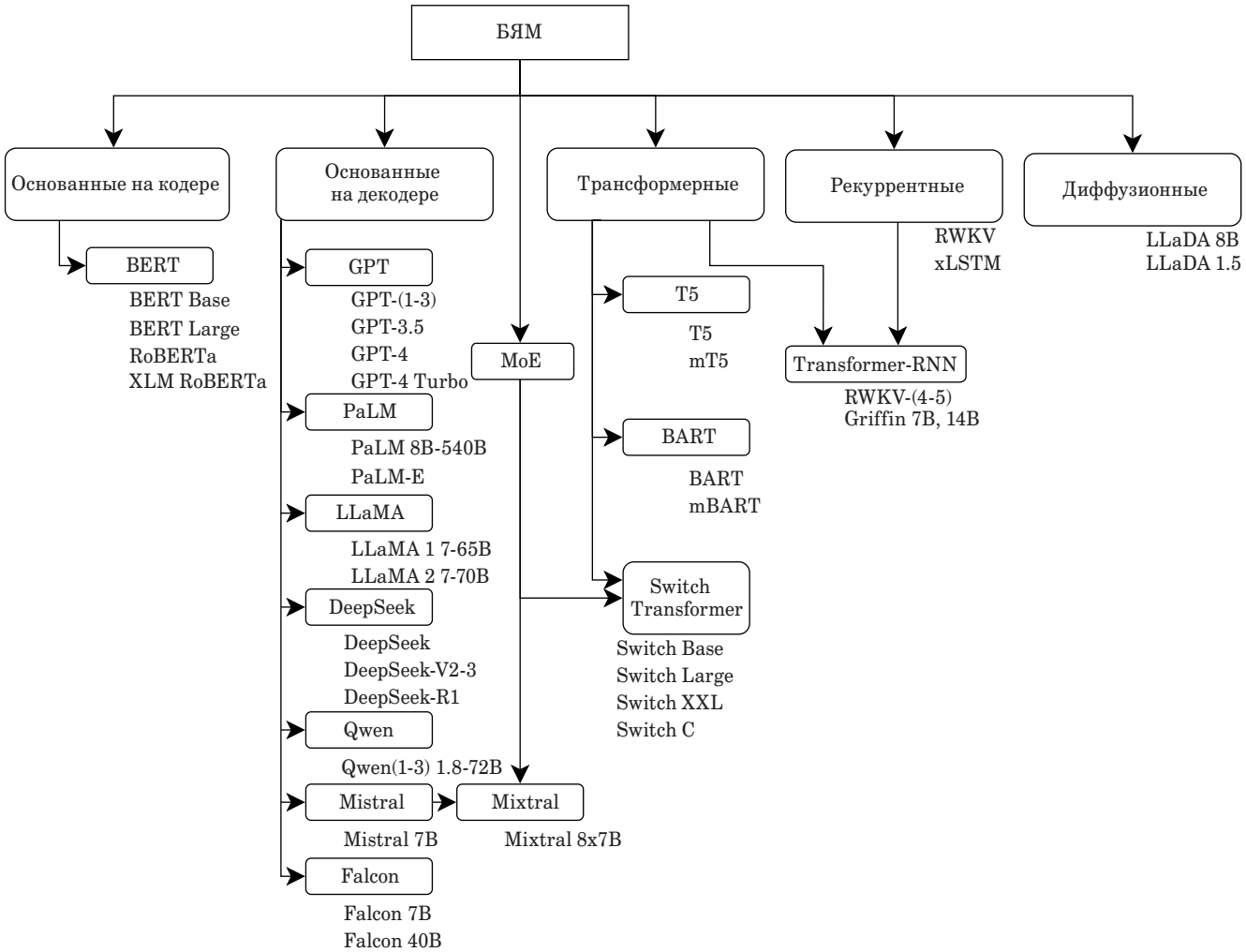
В моделях, состоящих только из кодера, каждый слой содержит механизм самовнимания и полносвязную сеть. Самовнимание является двунаправленным, т. е. при обработке токена модель может учитывать контекст как слева (предшествующие токены), так и справа (последующие токены). Эта архитектура применяется в задачах, связанных с векторными представлениями входной последовательности и не требующих генерации новых последовательностей, например при классификации текста, распознавании именованных сущностей, анализе тональности текста. Одной из моделей такого типа является модель BERT [12], которая в работе [13] была адаптирована для работы с русским языком (ruBERT). В работе [14] представлена RoBERTa, улучшенная модель BERT с особым процессом предобучения, обладающая повышенной устойчивостью к шумам. В работе [15] данная модель была адаптирована для задачи многоязычной обработки.

В моделях, состоящих только из декодера, самовнимание однонаправленное, т. е. при обработке токена модель обращает внимание только на текущий и предшествующий токены. Принцип работы такой модели авторегрессионный: она обрабатывает входную последовательность и на основе этого анализа генерирует следующий токен. Затем этот сгенерированный токен добавляется к последовательности, и процедура повторяется. Такая модель используется, прежде всего, для задач генерации текста. В качестве примера можно привести GPT [16], PaLM [17], LLaMA [18], Falcon [19] и Mistral [20].

Поскольку с ростом числа параметров модели увеличиваются как объем вычислений, так и требования к ресурсам, разработка методов оптимизации является актуальной задачей. Так,

к «оптимизированным» архитектурам можно отнести «смесь экспертов» (Mixture of Experts, MoE), принцип работы которой состоит в том, что вся модель делится на некоторое количество «экспертов» — отдельных подмоделей в общей архитектуре, которые в процессе обучения раздельно обрабатывают разные аспекты входных данных. При этом для отбора наиболее подходящих результатов используется дополнительная модель меньшего размера (распределитель). Дальнейшее развитие этой архитектуры привело к появлению «разреженных смесей экспертов» (sparse MoE), в которых распределитель не просто обрабатывает ответы, а предварительно выбирает наиболее подходящих экспертов для конкретной задачи, отключая остальные, что значительно снижает вычислительную нагрузку. Примером такого подхода служит модель Mixtral [21], являющаяся развитием декодерной модели Mistral. Mixtral использует архитектуру разреженной смеси экспертов, при которой для решения каждой задачи одновременно активируются только два из восьми экспертов. Другим примером является работа [22], в которой смесь экспертов совместно с архитектурой T5 использовалась для оптимизации трансформера Switch Transformer, что в итоге привело к значительному увеличению числа параметров модели, увеличению скорости обучения и снижению ресурсозатратности [22].

Рекуррентные БЯМ представляют собой класс архитектур, которые либо полностью основаны на рекуррентных нейронных сетях, либо сочетают в себе рекуррентные и трансформерные подходы. Известно, что трансформерные архитектуры легко распараллеливаются и допускают значительное увеличение объема обучаемых параметров, однако потребление памяти и вычислительная сложность растут квадратично относительно длины входной последовательности. В отличие от трансформеров, рекуррентные нейронные сети демонстрируют линейную зави-



■ ***Рис. 2.*** Основные архитектуры БЯМ
■ ***Fig. 2.*** Principal architectures of LLMs

симость затрат памяти и вычислительной сложности от длины последовательности, однако их параллелизация и масштабирование затруднены. Это послужило толчком к созданию новых архитектур, таких как Mamba [23], основанная на модели пространства состояний, или гибридных рекуррентно-трансформерных БЯМ, например RWKV [24], xLSTM [25] или Griffin [26].

Диффузионные БЯМ [27] представляют собой альтернативный подход к созданию генеративных нейросетевых моделей для восстановления маскированного и зашумленного текста. Они позволяют порождать текст без использования авторегрессии [28], а также моделировать двунаправленные зависимости между токенами. Языковые модели, основанные на диффузионных моделях, показывают превосходные результаты в задачах, связанных с выводом обратных заключений из заданных утверждений, однако исследования по их применению в области речевых технологий пока что не успели получить широкого распространения [29].

Представленные архитектуры БЯМ (рис. 2) иллюстрируют факт частого объединения различных архитектур в гибридные для уменьшения недостатков каждой из них. Также из схемы следует, что диффузионные языковые модели пока что получили меньшее распространение по сравнению с трансформерными и рекуррентными. Однако представляется, что высокая устойчивость диффузионных моделей к шумам обладает потенциалом в контексте задач малоресурсного языкового моделирования, когда обучающие данные невелики. В настоящее время БЯМ находят все большее применение для различных задач, в том числе для автоматического распознавания речи. В последующих разделах приведен обзор основных методов применения БЯМ для распознавания речи.

Применение БЯМ для переоценки списка гипотез

Для повышения точности распознавания часто используется метод переранжирования гипотез. Этот процесс начинается с того, что на первом этапе СРР генерирует не одну окончательную версию, а несколько наиболее вероятных вариантов или гипотез. Из них формируется список N лучших гипотез (N указывает на количество предложенных системой вариантов с наивысшими оценками). БЯМ вычисляет новые оценки для каждой гипотезы. Далее, на втором этапе, эти предварительно отобранные гипотезы подвергаются дополнительной оценке: БЯМ вычисляет новые, уточненные оценки для каждой гипотезы. В итоге исходная вероятностная оценка от СРР объединяется с оценкой, полученной от БЯМ, следующим образом:

$$w_{best} = \text{argmax}_{w \in W}[(1 - \lambda)\log P_{\text{СРР}}(w) + \lambda \log P_{\text{БЯМ}}(w)],$$

где w_{best} — выходная гипотеза с наибольшей вероятностью; w — гипотеза из списка лучших гипотез; λ — весовой коэффициент БЯМ; $P_{\text{СРР}}$, $P_{\text{БЯМ}}$ — вероятности гипотезы, полученные от СРР и БЯМ.

После этого выполняется переранжирование гипотез распознавания в соответствии с новыми вероятностными оценками и осуществляется выбор новой наилучшей гипотезы, т. е. гипотезы с наибольшей вероятностью. Аналогичным образом может выполняться переоценка не списка гипотез, а решетки слов, которая представляет собой граф гипотез с их вероятностными оценками.

Использование моделей языка на основе архитектуры BERT для переоценки гипотез распознавания подробно рассмотрено в работе [30]. Эксперименты на корпусе LibriSpeech показали, что применение BERT для повторной оценки 100 наиболее вероятных гипотез значительно повышает качество распознавания по сравнению с однонаправленными моделями.

В исследовании [31] задача переранжирования гипотез сформулирована как предсказание гипотезы с минимальным значением показателя неправильно распознанных слов (word error rate, WER) из списка N лучших гипотез. Именно эту идею авторы использовали при создании модели.

Работа [32] предлагает подход к переранжированию гипотез с использованием многомодальных БЯМ, объединяющих текстовые и речевые токены. Для получения речевых представлений используется HuBERT.

Важно отметить, что ограниченность списков лучших гипотез приводит к потере альтернативных вариантов. Именно поэтому в работе [33] предлагается переоценка всей решетки распознавания, что позволяет хранить больше гипотез в графовой структуре. БЯМ получает полное пространство гипотез и выдает единственную наилучшую. Сравнение показало эффективность метода на коротких фразах, но обнаружило ухудшение результатов для длинных фраз как при переоценке гипотез, так и при переоценке решетки. Авторы работы предполагают, что это связано с малыми размерами модели, а также с особенностью обучающих данных, представляющих собой небольшой набор спонтанной японской речи, поэтому длинных фраз в них немного. Также, предположительно, это может быть связано с ограничением длины контекста. Список

лучших гипотез или решетка для длинного предложения могут превысить длину контекста, которую способна обработать БЯМ. Кроме того, с увеличением длины фразы увеличивается возможное число ошибок, а ошибка в одном слове может приводить к ошибкам в других словах.

Применение БЯМ для коррекции ошибок

Большие языковые модели доказали свою высокую эффективность в исправлении орфографических и грамматических ошибок, что сделало их эффективным инструментом для корректировки гипотез, полученных в результате распознавания речи. В этой области выделяют два основных подхода: неограниченную и ограниченную коррекцию ошибок [34]. При неограниченной коррекции БЯМ генерирует полностью новую, исправленную гипотезу, опираясь на входной список лучших вариантов. Однако такой подход может приводить к избыточным изменениям, особенно когда обрабатывается лишь одна гипотеза.

В противоположность этому ограниченная коррекция ошибок требует от БЯМ выбора одной из предложенных гипотез, не допуская создания принципиально новых. Этот подход реализуется двумя способами: либо через селективный метод, при котором БЯМ непосредственно выбирает гипотезу из заданного списка, что по сути аналогично переоценке списка гипотез, либо через метод ближайшего соответствия, когда БЯМ сначала генерирует скорректированную гипотезу, а затем выбирает из исходного списка ту, что максимально близка (по показателю расстояние Левенштейна) к сгенерированному исправлению. Основное преимущество ограниченной коррекции заключается в снижении риска внесения новых ошибок, так как конечный результат всегда находится в пределах гипотез, изначально полученных от СРР. Тем не менее качество такой коррекции напрямую зависит от разнообразия и представительности этих исходных гипотез.

Различные БЯМ и стратегии их применения активно используются для оптимизации механизма коррекции ошибок. Так, в [35] подчеркивается важность стратегий расширения данных для обучения робастных моделей, а в [36] указывается эффективность метода ограничения коррекции решеткой распознавания. Современные крупные БЯМ, такие как ChatGPT, также активно применяются в задачах коррекции ошибок. Исследования [34, 37] показывают, что, хотя эти модели могут давать сопоставимые или лучшие результаты для определенных архитектур распознавания (например, Transformer-Transducer), их эффективность снижается для моделей, которые осуществляют сильную нормализацию текста, например Whisper. Это связано с тем, что нормализация уменьшает разнообразие гипотез, ограничивая пространство для коррекции. Одной из альтернатив решения этой проблемы является комбинация списков N лучших гипотез от различных систем распознавания, что позволяет использовать ошибки различных типов и улучшить общий результат коррекции [37].

С точки зрения вычислительных ресурсов рациональным является возможность коррекции только одной наилучшей гипотезы, как это предложено в [38] с применением многоязычной БЯМ Qwen1.5 7B. Этот подход позволил авторам снизить ресурсоемкость и заодно продемонстрировал эффективность переноса знаний между языками со схожей письменностью.

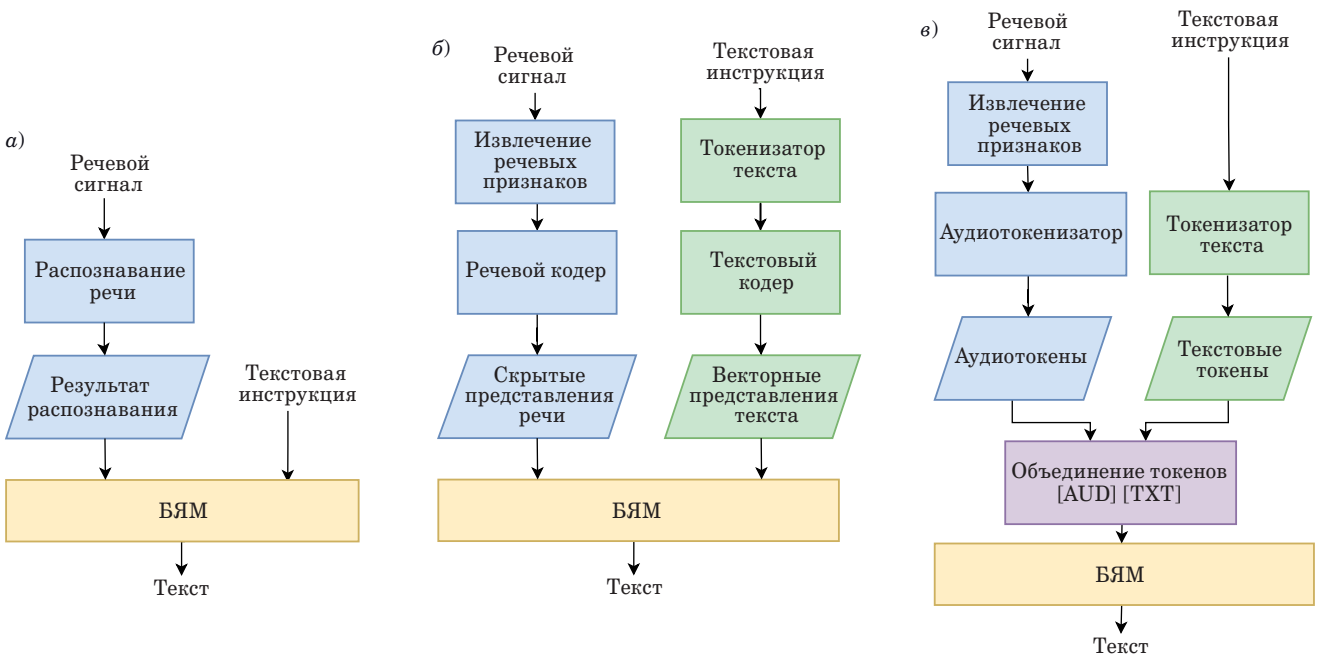
Серьезной проблемой при использовании БЯМ остается риск так называемых «галлюцинаций» — генерации недостоверной информации моделью, особенно если исходные ошибки системы распознавания минимальны. В таких случаях БЯМ может вносить ошибочные изменения, что в итоге приводит к увеличению значения показателя WER [39]. Для решения этой проблемы в [39] предлагается многопроходный метод коррекции, объединяющий результаты от различных систем распознавания и разных БЯМ с использованием алгоритма ROVER и компенсирующий недостатки отдельных моделей, что приводит к общему снижению риска галлюцинаций.

В целом использование БЯМ для коррекции ошибок часто дает лучшие результаты и происходит быстрее, чем переранжирование. Однако необходимо учитывать, что БЯМ могут генерировать синонимы, отличающиеся по звучанию от произнесенного слова [36], или вносить избыточные перефразирования [40], стремясь сделать текст более естественным. Контроль над этими изменениями может быть достигнут путем тщательного подбора запросов (prompt), как, например, сделано в [41] при обработке векторных представлений модели BART.

Объединение аудио и БЯМ

Аудио-БЯМ — это тип БЯМ, способный работать с аудиомодальностью за счет интеграции аудиоинформации в архитектуру БЯМ [42]. Основными способами интеграции аудиоинформации в БЯМ являются каскадная интеграция, интеграция на основе скрытых представлений и интеграция на основе аудиотокенов (рис. 3, *а–в*) [43].

Каскадная интеграция. Каскадная интеграция является самым простым методом ин-



■ ***Рис. 3.*** Схемы основных способов интеграции речевых и текстовых данных в аудио-БЯМ: *а* — каскадный; *б* — на основе скрытых представлений; *в* — на основе аудиотокенов

■ ***Fig. 3.*** Main approaches for integrating speech and text data into audio-LLMs: *а* — cascade; *б* — based on latent representations; *в* — based on audio tokens

теграции речевых данных в БЯМ. Речь вначале преобразуется в текст с помощью модуля автоматического распознавания речи, а затем полученный текст обрабатывается с помощью БЯМ [43]. Примером является модель, описанная в работе [44]: авторы использовали ряд фундаментальных моделей для обработки аудиоданных (в том числе Whisper для распознавания речи), при этом в качестве БЯМ-интерфейса служил ChatGPT. Преимуществом каскадного подхода является простота в реализации: во-первых, он позволяет использовать уже существующие, предварительно обученные СРР и БЯМ без необходимости дообучения и, во-вторых, не требует больших вычислительных ресурсов благодаря независимости этапов обработки данных.

Каскадная интеграция также использовалась в работе [29], в которой исследовалось применение диффузионной модели LLaDA 8B Instruct в нескольких сценариях: для улучшения результатов распознавания, полученных с помощью Whisper-LLaMA, для каскадного объединения с Whisper и в качестве самостоятельного декодера, заменяющего декодер Whisper. При каскадном объединении на вход модели подавалась не только текстовая транскрипция, но и векторные представления речевого сигнала, что привело к снижению WER на 12,3 % и улучшению результатов по сравнению с использованием Whisper-LLaMa. При этом введение только текста в качестве входных данных не позволило повысить точность распознавания. Также значительное повышение производительности продемонстрировало привлечение диффузионной модели в качестве декодера, однако данный подход не позволил достичь снижения WER.

Однако, несмотря на эти преимущества, каскадная интеграция имеет существенный недостаток: ошибки, допущенные системой СРР, неизбежно распространяются и влияют на последующую работу БЯМ. Кроме того, при таком подходе БЯМ не имеет прямого доступа к аудиопризнакам, что ограничивает ее возможности по учету нюансов речи.

Интеграция на основе скрытых представлений. Этот подход предполагает использование речевого кодера, который обрабатывает речевой сигнал и генерирует скрытые представления. Эти представления затем напрямую подаются в БЯМ, минуя промежуточный этап преобразования в текст [43]. Речевой кодер может быть обучен с нуля, или же в нем используются предварительно обученные модели, такие как HuBERT [45]. Основная сложность здесь заключается в согласовании длины последовательностей, поскольку речевые признаки обычно значительно длиннее текстовых токенов.

Первые эксперименты по такому объединению речевых и текстовых данных на основе аудиотокенов предложены в исследовании [46], где в базовую СРР была интегрирована модель BERT для работы с путунхуа. Акустические при-

знаки выравнивались с текстовыми токенами, при этом каждому слову в тексте соответствовал сегмент аудиофреймов. Далее аудиофреймы преобразовывались аудиокодером в векторные представления той же размерности, что и в BERT. Тем не менее авторам не удалось превзойти результаты базовой модели DNN-HMM.

Более поздние работы демонстрируют значительный прогресс. Например, в [47] представлена мультизадачная многоязычная модель SLM, которая обрабатывает как текстовые, так и речевые данные. Она основана на предобученной универсальной речевой модели Google USM [48] и различных моделях T5. В этой работе веса исходных моделей были заморожены, и был обучен небольшой адаптер (1 % параметров), преобразующий речевые векторные представления (эмбеддинги) в текстовые. Затем эти представления подаются в БЯМ вместе с внешними текстовыми представлениями. В работе отмечается, что, несмотря на улучшение точности работы, модель иногда может порождать галлюцинации. В статье [49] также был предложен специализированный конвертер, согласующий аудиокодер с БЯМ, что позволяет преобразовывать речевые представления в совместимое векторное пространство. Эта модель, обученная по стратегии «обучаемый аудиокодер + обучаемый конвертер + фиксированная БЯМ», показала высокие результаты во многих задачах, включая автоматическое распознавание речи и ведение диалога. Интеграция на основе скрытых представлений позволяет снизить уровень ошибки, поскольку БЯМ учитывает аудиопризнаки, которые были бы утрачены при использовании промежуточного преобразования аудио в текст. Тем не менее многие из этих подходов ориентированы на дообучение адаптера под конкретную задачу, и они обычно имеют более высокие требования к вычислительным ресурсам и объему данных для обучения.

В работе нашего коллектива [50] рассматривалось применение модели W2V2-BERT v2, основанной на объединении аудиокодера и языковой модели BERT, для малоресурсного распознавания речи на карельском языке. Продемонстрированные результаты значительно превзошли аналоги на основе дообучения только аудиокодера, аудиокодера со статистической языковой моделью и Whisper. Авторы еще одного исследования, посвященного распознаванию речи на малоресурсных языках, в частности тайском и вьетнамском [51], представили интеграцию БЯМ Qwen2.5 и Gemma3 и аудиокодера Whisper.

Интеграция на основе аудиотокенов. При этом подходе речевой сигнал преобразуется в дискретные единицы, которые затем подаются на вход БЯМ аналогично текстовым токенам. В работе [52] предложена модель SpeechGPT, предназначенная для распознавания и генерации речи. Она состоит из речевого кодера в дискретные токены (использующего HuBERT), расширенной БЯМ LLaMA, в словарь которой добавлены речевые токены, и речевого декодера, который преобразует речевые токены обратно в аудиосигнал. Дискретные токены также используются в мультимодальной модели AudioPaLM [53], которая обрабатывает и генерирует текст и речь, выполняя распознавание, перевод и голосовой перевод. Авторы исследовали модели на основе PaLM [17] и PaLM 2 [54] с 8 млрд параметров, предобученные только на текстовых данных. В работе [55] предлагаются так называемые дискретные речевые единицы, генерируемые из скрытого представления речевого кодера путем кластеризации, которые преобразуются речевым адаптером в векторные представления БЯМ, которые конкатенируются с текстовым запросом. Преимущество такого подхода — в применимости к задачам не только распознавания, но и синтеза речи. К недостаткам относятся высокие вычислительные требования и меньшая точность по сравнению с интеграцией на основе скрытых представлений [43].

Стоит отметить, что для распознавания речи предпочтительна интеграция на основе скрытых представлений, тогда как для синтеза — на основе токенов. Примером комбинированного подхода является LauraGPT [56], которая кодирует входное аудио в непрерывные представления, а выходные данные генерирует из дискретных кодеков. В современных исследованиях эти подходы получают дальнейшее развитие с фокусом на оптимизации для различных сценариев. Например, в работе [57] представлен подход к распознаванию речи с переключением кодов (китайский-английский) на основе БЯМ Qwen2 7B с MoE и токеном прерывания, что обеспечивает согласование между генерацией текста и CPP.

Контекстуализация и контекстное обучение

Одной из важных особенностей БЯМ является способность учитывать контекст, т. е. извлекать контекстную информацию из входной последовательности при генерации выходной последовательности. Достаточно большое число исследований посвящено контекстуализации для улучшения текста, генерируемого БЯМ. В качестве примера можно привести работу [58], авторы которой предложили модель, названную Speech LLaMA. Модель состоит из двух компонентов:

аудиокодера, преобразующего речь в скрытые представления, и БЯМ-декодера (LLaMA 7B), порождающего текст на основе аудио и контекста (например, заголовка и описания видео). Эксперименты показали снижение значения WER на 6 %.

Исследования в области контекстуализации привели к возникновению контекстного обучения (in-context learning), которое чаще всего применяется для переоценки или корректировки гипотез распознавания в задачах распознавания речи. При этом, помимо гипотезы распознавания, которую необходимо откорректировать, на вход БЯМ подаются примеры исправлений, а также соответствующие запросы, что можно формализовать следующим образом [59]:

$$y = f(I, (x_1, y_1), (x_2, y_2), \ldots, (x_k, y_k), x),$$

где y — исправленный вывод; $f(\cdot)$ — функция контекстного исправления гипотез распознавания, реализуемая БЯМ; I — запрос для БЯМ; x — гипотеза распознавания, подлежащая корректировке; $(x_i, y_i)_{i=1}^{k}$ — примеры исправления ошибок, где k — число примеров.

Например, контекстное обучение для корректировки результатов распознавания использовано в работе [59], посвященной применению различных версий GPT-3.5 и GPT-4 для работы с корпусом LibriSpeech и корпусом китайской речи Aishell-1. На вход БЯМ подавался результат распознавания и соответствующие запросы для исправления потенциальных ошибок. В работе рассмотрено несколько стратегий: введение запросов с варьированием степени детализации, обучение на одном, двух и трех примерах, несколько попыток коррекции с выбором результата с наименьшим WER. Тем не менее авторам не удалось снизить WER за счет применения БЯМ. Более подробные запросы, а также предоставление большего числа примеров повышали точность, однако получаемый WER был все равно выше исходного. Даже при разрешении нескольких попыток исправления (до пяти) с выбором лучшего результата исправления БЯМ все равно вносили больше ошибок.

В работе [60] выполнено сравнение контекстного обучения, дообучения и низкоранговой адаптации для исправления ошибок распознавания. Были рассмотрены T5, LLaMA, GPT-3.5. Не имевшие предварительных примеров БЯМ с малым количеством параметров не дали заметных улучшений при распознавании с использованием Whisper, но использование БЯМ с большим количеством параметров и представлением контекста от WavLM позволило значительно улучшить точность, особенно в зашумленном и малоресурсных контекстах.

В работе [61] предложен метод контекстного обучения, названный авторами задачно-ориентированными запросами (task-activating prompting). Отличие от традиционного контекстного обучения состоит в том, что контекстное обучение происходит за один этап, состоящий из запроса, примеров и текста, подаваемого на вход. Задачно-ориентированные запросы — это многоэтапный процесс, состоящий из последовательности вопросов и ответов. Например, модель сначала спрашивают, знает ли она, что такое автоматическое распознавание речи, затем просят привести пример исправления ошибок и только после этого дают конкретные данные для обработки. В результате применения этого метода авторы смогли добиться уменьшения метрики WER на 31–38 %.

Обсуждение

Проведен сравнительный анализ методов применения БЯМ для распознавания речи по относительному сокращению WER (таблица). В части представленных работ оценка распознавания проводилась по показателю количества неправильно распознанных символов (character error rate, CER).

Из таблицы видно, что применение БЯМ для переранжирования гипотез демонстрирует стабильное улучшение результатов распознавания, но в то же время во многих случаях наблюдаемое улучшение лишь незначительно превосходит показатели эталонных систем, не использующих БЯМ. Применение БЯМ для коррекции ошибок часто демонстрирует улучшение результатов, при этом задача выполняется несколько быстрее, чем при переранжировании. Генеративный характер БЯМ позволяет им порождать исправленный текст напрямую, минуя фазу многоэтапного ранжирования, что обеспечивает существенное повышение скорости обработки. Однако, несмотря на очевидные преимущества в эффективности и скорости, серьезной проблемой в данном случае остается риск галлюцинаций, в принципе присущий генеративным моделям. В контексте коррекции ошибок это может означать не только неспособность исправить существующую ошибку, но и внесение новой, некорректной информации в текст. Работы с применением контекстуализации еще больше разнятся по результатам, демонстрируя как значительные улучшения по сравнению с базовыми моделями, так и ухудшения. Этот разброс может быть связан со сложностью применения БЯМ на основе декодировщиков и необходимостью тщательного подбора запросов, однако при условии использования неглубоких моделей и детальной контекстуализации дан-

■ Сравнительный анализ методов применения БЯМ для распознавания речи
■ Comparative analysis of LLM application methods for ASR tasks

Ссылка	Архитектура CPP	Архитектура БЯМ	Речевой корпус	Относительное сокращение WER, %
Переранжирование гипотез распознавания				
[30]	Listen, Attend and Spell (LAS)	BERT	LibriSpeech Clean	22,18
			LibriSpeech Other	14,73
[31]	TDNN/HMM	BERT	AMI	4,39
[32]	Whisper large v2	330M, аналогичная OPT	LibriSpeech Clean/Other	17,70/12,92
		7B, аналогичная Llama		18,14/14,79
Коррекция ошибок распознавания				
[33]	FSMN + 3-граммная языковая модель	BART	Собственный корпус путунхуа	CER: 21,85
[36]	Conformer-Transducer	T5	LibriSpeech Clean	12,15
			LibriSpeech Other	11,19
[34]	Conformer-Transducer	ChatGPT	LibriSpeech Other	10,14
	Whisper			5,41
[37]	Conformer-Transducer	GPT-4	LibriSpeech Other	31,59
	Whisper Small.en			–2,83
[38]	MMS	Qwen1.5	SPREDS-U1 (20 языков) (ast-astrec.nict.go.jp/en/release/SPREDS-U1)	CER: 70,16 (англ.)/ 31,70 (рус.)
	OWSM v3.1			CER: 34,65 (англ.)/ 46,45 (рус.)
	Whisper v3			CER: 44,19 (англ.) / 5,95 (рус.)
[39]	Комбинация моделей Whisper (различных версий), MMS, OWSM v3.1	ELYZA 7B, Qwen1.5 7B	SPREDS-U1-ja (японский)	39,81 (одной моделью)
				45,24
			CSJ (японский) [62]	6,67
Использование контекстного обучения				
[59]	Гибридная архитектура CTC/attention (предобученные веса от Wenet)	GPT-3.5 (разные версии), GPT-4	LibriSpeech Clean	–374,62
			LibriSpeech Other	–31,01
			Aishell-1 (китайский)	–21,99
[60]	WavLM, Whisper	T5	Различные корпуса, например WSJ	40
		LLaMA		51,11
Объединение БЯМ с аудиокодировщиком				
[51]	Whisper large-V3	Qwen2.5	Тайский	–22,14
			Вьетнамский	12,52
		Gemma3	Тайский	8,56
			Вьетнамский	11,63
[29]	Whisper	LLaDA 8B Instruct	LibriSpeech Other	12,3
[57]	Whisper-large-V3	Qwen2 7B с MoE и IDIT	Китайский с переключением на английский	19,83

ный подход может демонстрировать значительно лучшие результаты, чем прочие. Применение аудио-БЯМ к материалу малоресурсных языков показывает как улучшение, так и ухудшение результатов. Стоит отметить, однако, что использование многоязычной БЯМ, оснащенной механизмом «смеси экспертов», позволило существенно снизить значение показателя WER при распознавании китайской речи с переключением на английский. Этот факт указывает на перспективность дальнейших исследований по применению БЯМ для распознавания речи с переключением кодов в малоресурсных языках.

Проведенный анализ работ показывает, что некоторые архитектуры БЯМ, в частности рекуррентные и диффузионные, пока что не получили широкого распространения в контексте задач по распознаванию речи (несмотря на устойчивость последних к шумам). В целом наблюдается тенденция к использованию моделей, основанных на архитектуре декодера. Эта тенденция в свою очередь связана с рядом проблем, включая склонность этих моделей к галлюцинациям и их направленность на форматирование текста в соответствии с письменными нормами. Последнее обстоятельство может существенно затруднять точную передачу сказанного «слово в слово».

Кроме того, БЯМ могут использоваться совместно с СРР не только для повышения точности распознавания, но также для предобработки данных и постобработки результатов распознавания. В частности, способность БЯМ выполнять генерацию текстов может использоваться для аугментации текстовых данных для последующего обучения языковых моделей, что может быть особенно полезным при создании СРР для малоресурсных языков и речи с переключением кодов [63]. Постобработка результатов распознавания с помощью БЯМ может заключаться в резюмировании распознанного текста [64], переводе на другой язык [65], генерации ответов на речевые запросы пользователя [66], диаризации речи дикторов [67].

Также не трудно заметить, что различные методы применения БЯМ для распознавания речи в российских исследованиях представлены мало, хотя имеются подобные работы по применению акустических моделей на основе трансформера для распознавания русской речи [68] и моделей с интеграцией на основе латентных представлений аудиокодера и языковой модели BERT для распознавания карельской речи [50]. В контексте постобработки результатов распознавания с помощью БЯМ российскими учеными также рассматривалась задача генерации ответов на речевые запросы пользователя, например, в работе [69] описывается применение BERT и GPT-2 для этих задач.

Заключение

В настоящей статье систематизированы и обобщены существующие способы применения больших языковых моделей в СРР. Особое внимание уделено использованию БЯМ для переоценки списка гипотез, коррекции ошибок, а также различным способам интеграции аудиоданных в БЯМ.

Анализ показал, что БЯМ действительно обладают потенциалом к значительному улучшению результатов распознавания — за счет эффективных механизмов переранжирования гипотез или прямой коррекции ошибок. Однако наблюдаемое улучшение не всегда является кардинальным по сравнению с эталонными системами, не использующими БЯМ. Кроме того, применение генеративных БЯМ сопряжено с такими проблемами, как галлюцинации и избыточные коррекции при порождении письменного представления текста, что безусловно влияет на точность распознавания устной речи. Тем не менее разнообразие методов интеграции аудиоданных и БЯМ — от каскадных до основанных на скрытых представлениях и аудиотокенах, при том, что каждый из них имеет свои преимущества и недостатки — является самим по себе мощным инструментом для дальнейшего повышения эффективности БЯМ в задачах распознавания речи.

В целом, несмотря на уже достигнутые успехи, использование БЯМ в распознавании речи все еще находится на стадии активного развития. Применение новейших архитектур БЯМ, в частности рекуррентных и диффузионных моделей, которые обладают повышенной устойчивостью к шуму, представляется весьма перспективным направлением. Дальнейшие исследования могут быть связаны с улучшением методов, позволяющих эффективно контролировать генерацию БЯМ для минимизации галлюцинаций и обеспечения точности передачи устной речи, а также с исследованием потенциала БЯМ в контексте малоресурсного распознавания и переключения кодов.

Финансовая поддержка

Данное исследование выполнено в рамках бюджетной темы СПб ФИЦ РАН (№ FFZF-2025-0003).

Литература

1. **Zhao W. X., Zhou K., Li J., Tang T., Wang X., Hou Y., Min Y., Zhang B., Zhang J., Dong Z., Du Y., Yang C., Chen Y., Chen Z., Jiang J., Ren R., Li Y., Tang X., Liu Z., Liu P., Nie J. Y., Wen J. R.** A Survey of large language models. *arXiv preprint*, 2023. arXiv:2303.18223. doi:10.48550/arXiv.2303.18223
2. **Minaee Sh., Mikolov T., Nikzad N., Chenaghlu M. A., Socher R., Amatriain X., Gao J.** Large language models: A survey. *arXiv preprint*, 2024. arXiv:2402.06196. doi:10.48550/arXiv.2402.06196
3. **Wang F., Zhang Z., Zhang X., Wu Z., Mo T., Lu Q., Wang W., Li R., Xu J., Tang X., He Q., Ma Y., Huang M., Wang S.** A comprehensive survey of small language models in the era of large language models: Techniques, enhancements, applications, collaboration with LLMs, and trustworthiness. *arXiv preprint*, 2024. arXiv:2411.03350. doi:10.48550/arXiv.2411.03350
4. **Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser Ł., Polosukhin I.** Attention is all you need. *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS-2017)*, 2017, pp. 6000–6010. doi:10.48550/arXiv.1706.03762
5. **Капуста К. Л., Кипяткова И. С., Кагиров И. А.** Аналитический обзор интегральных моделей и стратегий распознавания речи на основе архитектуры трансформер. *Информационно-управляющие системы*, 2024, № 5, с. 2–15. doi:10.31799/1684-8853-2024-5-2-15, EDN: MWTGXE
6. **Hwang S., Lahoti A., Puduppully R., Dao T., Gu A.** Hydra: Bidirectional state space models through generalized matrix mixers. *Proceedings of the 38th Annual Conference on Neural Information Processing Systems (NIPS-2024)*, pp. 110876–110908. doi:10.48550/arXiv.2407.09941
7. **Xu W., Hu W., Wu F., Sengamedu S.** DeTiME: Diffusion-enhanced topic modeling using encoder-decoder based LLM. *Findings of the Association for Computational Linguistics (EMNLP-2023)*, 2023, pp. 9040–9057. doi:10.18653/v1/2023.findings-emnlp.606
8. **Lewis M., Liu Y., Goyal N., Ghazvininejad M., Mohamed A., Levy O., Stoyanov V., Zettlemoyer L.** BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 7871–7880. doi:10.18653/v1/2020.acl-main.703
9. **Liu Y., Gu J., Goyal N., Li X., Edunov S., Ghazvininejad M., Lewis M., Zettlemoyer L.** Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 2020, vol. 8, pp. 726–742. doi:10.1162/tacl_a_00343
10. **Raffel C., Shazeer N., Roberts A., Lee K., Narang Sh., Matena M., Zhou Y., Li W., Liu P. J.** Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 2020, vol. 21, iss. 1, pp. 5485–5551. doi:10.48550/arXiv.1910.10683
11. **Xue L., Constant N., Roberts A., Kale M., Al-Rfou R., Siddhant A., Barua A., Raffel C.** mT5: A massively multilingual pre-trained text-to-text transformer. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-2021): Human Language Technologies*, 2021, pp. 483–498. doi:10.18653/v1/2021.naacl-main.41
12. **Devlin J., Chang M. W., Lee K., Toutanova K.** BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-2021): Human Language Technologies*, 2021, pp. 4171–4186. doi:10.18653/v1/N19-1423
13. **Kuratov Yu., Arkhipov M.** Adaptation of deep bidirectional multilingual transformers for Russian language. *arXiv preprint*, 2019. arXiv:1905.07213. doi:10.48550/arXiv.1905.07213
14. **Liu Y., Ott M., Goyal N., Du J., Joshi M., Chen D., Levy O., Lewis M., Zettlemoyer L., Stoyanov V.** RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint*, 2019. arXiv:1907.11692. doi:10.48550/arXiv.1907.11692
15. **Conneau A., Khandelwal K., Goyal N., Chaudhary V., Wenzek G., Guzmán F., Grave E., Ott M., Zettlemoyer L., Stoyanov V.** Unsupervised cross-lingual representation learning at scale. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL-2020)*, 2020, pp. 8440–8451. doi:10.18653/v1/2020.acl-main.747
16. **Radford A., Narasimhan K., Salimans T., Sutskever I.** Improving language understanding by generative pre-training. *OpenAI Blog*, 2018. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf (дата обращения: 14.05.2025).
17. **Chowdhery A., Narang S., Devlin J., Bosma M., et al.** PaLM: Scaling language modeling with pathways. *The Journal of Machine Learning Research*, 2023, vol. 24, iss. 1, pp. 11324–11436. doi:10.48550/arXiv.2204.02311
18. **Touvron H., Lavril T., Izacard G., Martinet X., Lachaux M. A., Lacroix T., Rozière B., Goyal N., Hambro E., Azhar F., Rodriguez A., Joulin A., Grave E., Lample G.** Llama: Open and efficient foundation language models. *arXiv preprint*, 2023. arXiv:2302.13971. doi:10.48550/arXiv.2302.13971
19. **Almazrouei E., Alobeidli H., Alshamsi A., Cappelli A., Cojocaru R., Debbah M., Goffinet E., Heslow D., Launay J., Malartic Q., Mazotta D., Noune B., Pannier B., Penedo G.** The Falcon series of open language models. *arXiv preprint*, 2023. arXiv:2311.16867. doi:10.48550/arXiv.2311.16867

20. **Jiang D., Wu B., Chen C., Li R., Chen G., Sun Y., Kong X., Li L.** From clip to dino: Visual encoders shout in multi-modal large language models. *arXiv preprint*, 2023. arXiv:2310.08825. doi:10.48550/arXiv.2310.08825
21. **Jiang A. Q., Sablayrolles A., Roux A., Mensch A., Savary B., Bamford C., Chaplot D. S., de las Casas D., Hanna E. B., Bressand F., Lengyel G., Bour G., Lample G., Lavaud L. R., Saulnier L., Lachaux M.-A., Stock P., Subramanian S., Yang S., Antoniak S., Le Scao T., Gervet T., Lavril T., Wang T., Lacroix T., El Sayed W.** Mixtral of experts. *arXiv preprint*, 2024. arXiv:2401.04088. doi:10.48550/arXiv.2401.04088
22. **Fedus W., Zoph B., Shazeer N.** Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *The Journal of Machine Learning Research*, 2022, vol. 23, iss. 1, pp. 1–39. doi:10.48550/arXiv.2101.03961
23. **Gu A., Dao T.** Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint*, 2023. arXiv:2312.00752. doi:10.48550/arXiv.2312.00752
24. **Peng B., Alcaide E., Anthony Q., Albalak A., Arcadinho S., Biderman S., Cao H., Cheng X., Chung M., Grella M., Kiran G. K., He X., Hou H., Lin J., Kazienko P., Kocon J., Kong J., Koptyra B., Lau H., Mantri K. S. I., Mom F., Saito A., Song G., Tang X., Wang B., Wind J. S., Wozniak S., Zhang R., Zhang Z., Zhao Q., Zhou P., Zhou Q., Zhu J., Zhu R.-J.** RWKV: Reinventing RNNs for the transformer era. *Findings of the Association for Computational Linguistics (EMNLP-2023)*, 2023, pp. 14048–14077. doi:10.18653/v1/2023.findings-emnlp.936
25. **Beck M., Pöppel K., Spanring M., Auer A., Prudnikova O., Kopp M., Klambauer G., Brandstetter J., Hochreiter S.** xLSTM: Extended long short-term memory. *Advances in Neural Information Processing Systems*, 2025, vol. 37, pp. 107547–107603. doi:10.48550/arXiv.2405.04517
26. **De S., McLeish T., Botev A., Gu A., Dao T., Goyal N.** Griffin: Mixing gated linear recurrences with local attention for efficient language models. *arXiv preprint*, 2024. arXiv:2402.19427. doi:10.48550/arXiv.2402.19427
27. **Li X., Thickstun J., Gulrajani I., Liang P. S., Hashimoto T. B.** Diffusion-LM improves controllable text generation. *Advances in Neural Information Processing Systems*, 2022, vol. 35, pp. 4328–4343. doi:10.48550/arXiv.2205.14217
28. **Nie S., Zhu F., You Z., Zhang X., Ou J., Hu J., Li C.** Large language diffusion models. *Proceedings of Workshop on Deep Generative Model in Machine Learning: Theory, Principle and Efficacy (ICLR-2025)*, 2025. doi:10.48550/arXiv.2501.11720
29. **Wang M., Liu Zh., Jin Z., Sun G., Zhang Ch., Woodland P. C.** Audio-conditioned diffusion LLMs for ASR and deliberation processing. *arXiv preprint*, 2025. arXiv:2509.16622. doi:10.48550/arXiv.2509.16622
30. **Shin J., Lee Y., Jung K.** Effective sentence scoring method using BERT for speech recognition. *Proceedings of Machine Learning Research*, 2019, pp. 1081–1093. doi:10.48550/arXiv.1910.09932
31. **Chiu S. H., Chen B.** Innovative BERT-based reranking language models for speech recognition. *IEEE Spoken Language Technology Workshop (SLT-2021)*, 2021, pp. 266–271. doi:10.1109/SLT48900.2021.9383578
32. **Shivakumar P. G., Kolehmainen J., Gourav A., Gu Y., Gandhe A., Rastrow A., Bulyko I.** Speech recognition rescoring with large speech-text foundation models. *Proceedings of 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-2025)*, 2025, pp. 1–5. doi:10.1109/ICASSP48485.2025.10494321
33. **Li S., Ko Y., Ito A.** LLM as decoder: Investigating lattice-based speech recognition hypotheses rescoring using LLM. *Proceedings of 2024 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC-2024)*, 2024, pp. 1–5. doi:10.1109/APSIPAASC58517.2024.10373582
34. **Ma R., Qian M., Manakul P., Gales M., Knill K.** Can generative large language models perform ASR error correction? *arXiv preprint*, 2023. arXiv:2307.04172. doi:10.48550/arXiv.2307.04172
35. **Zhao Y., Yang X., Wang J., Gao Y., Yan C., Zhou Y.** BART based semantic correction for Mandarin automatic speech recognition system. *Proceedings of the 22nd Annual Conference of the International Speech Communication Association, Interspeech 2021*, 2021, pp. 2017–2021. doi:10.21437/Interspeech.2021-1023
36. **Ma R., Gales M. J., Knill K. M., Qian M.** N-best T5: Robust ASR error correction using multiple input hypotheses and constrained decoding space. *Proceedings of the 24th Annual Conference of the International Speech Communication Association, Interspeech 2023*, 2023, pp. 3267–3271. doi:10.21437/Interspeech.2023-2189
37. **Ma R., Qian M., Gales M., Knill K.** ASR error correction using large language models. *IEEE Transactions on Audio, Speech and Language Processing*, 2025, vol. 33, pp. 1389–1401. doi:10.1109/TASLPRO.2025.3551083
38. **Li S., Chen C., Kwok C. Y., Chu C., Cheng E. S., Kawai H.** Investigating ASR error correction with large language model and multilingual 1-best hypotheses. *Proceedings of the 25th Annual Conference of the International Speech Communication Association, Interspeech 2024*, 2024, pp. 1315–1319. doi:10.21437/Interspeech.2024-368
39. **Ko Y., Li S., Yang C. H. H., Kawahara T.** Benchmarking Japanese speech recognition on ASR-LLM setups with multi-pass augmented generative error correction. *arXiv preprint*, 2024. arXiv:2408.16180. doi:10.48550/arXiv.2408.16180
40. **Wu H., Wang W., Wan Y., Jiao W., Lyu M.** ChatGPT or Grammarly? Evaluating ChatGPT on grammatical

error correction benchmark. *arXiv preprint*, 2023. arXiv:2303.13648. doi:10.48550/arXiv.2303.13648

41. **Mirbeygi M., Beigy H.** Prompt guided diffusion for controllable text generation. *Proceedings of the Tenth Workshop on Noisy and User-generated Text*, 2025, pp. 78–84. doi:10.18653/v1/2025.wnut-1.9
42. **Li Y., Wang X., Cao S., Zhang Y., Ma L., Xie L.** A transcription prompt-based efficient audio large language model for robust speech recognition. *Proceedings of the 25th Annual Conference of the International Speech Communication Association, Interspeech 2024*, 2024, pp. 1905–1909. doi:10.21437/Interspeech.2024-968
43. **Yang Z., Chen X., Zhang H., Li Y., Wang Y., Yu D.** When large language models meet speech: A survey on integration approaches. *arXiv preprint*, 2025. arXiv:2502.19548. doi:10.48550/arXiv.2502.19548
44. **Huang R., Li M., Yang D., Shi J., Chang X., Ye Z., Watanabe S.** AudioGPT: Understanding and generating speech, music, sound, and talking head. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024, vol. 38, no. 21, pp. 23802–23804. doi:10.1609/aaai.v38i21.30570
45. **Hsu W. N., Bolte B., Tsai Y. H. H., Lakhotia K., Salakhutdinov R., Mohamed A.** Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2021, vol. 29, pp. 3451–3460. doi:10.1109/TASLP.2021.31222
46. **Huang W. C., Chen Z., Chuang P. Y., Harwath D., Glass J.** Speech recognition by simply fine-tuning BERT. *arXiv preprint*, 2021. arXiv:2102.00291. doi:10.48550/arXiv.2102.00291
47. **Wang M., Han W., Shafran I., Wu Z., Chiu C.-C., Cao Y., Wang Y., Chen N., Zhang Y., Soltau H., Rubenstein P., Zilka L., Yu D., Meng Z., Pundak G., Siddhartha N., Schalkwyk J., Wu Y.** SLM: Bridge the thin gap between speech and text foundation models. *Proceedings of 2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU-2023)*, 2023, pp. 1–8. doi:10.1109/ASRU57964.2023.10389703
48. **Zhang Y., Qin J., Park D. S., Han W., Chiu C. C., Pang R., Le Q. V., Wu Y.** Google USM: Scaling automatic speech recognition beyond 100 languages. *arXiv preprint*, 2023. arXiv:2303.01037. doi:10.48550/arXiv.2303.01037
49. **Bai Y., Chen J., Chen J., Chen W., Chen Z., Ding C., Dong L., Dong Q., Du Y., Gao K., Gao L., Guo Y., Han M., Han T., Hu W., Hu X., Hu Y., Hua D., Huang L., Huang M., Huang Y., Jin J., Kong F., Lan Z., Li T., Li X., Li Z., Lin Z., Liu R., Liu S., Lu L., Lu Y., Ma J., Ma S., Pei Y., Shen C., Tan T., Tian X., Tu M., Wang B., Wang H., Wang Y., Wang Y., Xia H., Xia R., Xie S., Xu H., Yang M., Zhang B., Zhang J., Zhang W., Zhang Y., Zhang Y., Zheng Y., Zou M.** SEED-ASR: Understanding diverse speech and contexts with LLM-based speech recognition. *arXiv preprint*, 2024. arXiv:2407.04675. doi:10.48550/arXiv.2407.04675
50. **Кипяткова И. С., Кагиров И. А., Долгушин М. Д.** Применение предварительно обученных многоязычных моделей для распознавания карельской речи. *Информатика и автоматизация*, 2025, № 24(2), с. 604–630. doi:10.15622/ia.24.2.9
51. **Nguyen T., Hoang L. V., Tran H. D.** Qwen vs. Gemma integration with Whisper: A comparative study in multilingual SpeechLLM systems. *Proceedings of Workshop on Multilingual Conversational Speech Language Model (MLC-SLM)*, 2025. doi:10.21437/MLCSLM.2025-10
52. **Zhang D., Li S., Zhang X., Zhan J., Wang P., Zhou Y., Qiu X.** SpeechGPT: Empowering large language models with intrinsic cross-modal conversational abilities. *Findings of the Association for Computational Linguistics (EMNLP-2023)*, 2023, pp. 15757–15773. doi:10.18653/v1/2023.findings-emnlp.1055
53. **Rubenstein P. K., Asawaroengchai C., Nguyen D. D., Bapna A., Borsos Z., Chaumont Quitry de F., Chen P., El Badawy D., Han W., Kharitonov E., Muckenhirn H., Padfield D., Qin J., Rozenberg D., Sainath T., Schalkwyk J., Sharifi M., Ramanovich T. M., Tagliasacchi M., Tudor A., Velimirović M., Vincent D., Yu J., Wang Y., Zayats V., Zeghidour N., Zhang Y., Zhang Zh., Zilka L., Frank Ch.** AudioPaLM: A large language model that can speak and listen. *arXiv preprint*, 2023. arXiv:2306.12925. doi:10.48550/arXiv.2306.12925
54. **Anil R., Dai A. M., Firat O., Johnson M., Lepikhin D., et al.** PaLM 2 technical report. *arXiv preprint*, 2023. arXiv:2305.10403. doi:10.48550/arXiv.2305.10403
55. **Shon S., Yang C. H. H., Lee H., Kim J., Kim S., Kim T., Lee H. Y.** DiscreteSLU: A large language model with self-supervised discrete speech units for spoken language understanding. *Proceedings of the 25th Annual Conference of the International Speech Communication Association, Interspeech 2024*, 2024, pp. 4154–4158. doi:10.21437/Interspeech.2024-1306
56. **Du Z., Wang J., Chen Q., Chu Y., Gao Z., Li Z., Zhang S.** LauraGPT: Listen, attend, understand, and regenerate audio with GPT. *arXiv preprint*, 2023. arXiv:2310.04673. doi:10.48550/arXiv.2310.04673
57. **Zhang F., Geng W., Huang H., Shan Y., Yi C., Qu H.** Boosting code-switching ASR with mixture of experts enhanced speech-conditioned LLM. *Proceedings of 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-2025)*, 2025, pp. 1–5. doi:10.1109/ICASSP49660.2025.10890030
58. **Lakomkin E., Wu C., Fathullah Y., Kalinli O., Seltzer M. L., Fuegen C.** End-to-end speech recognition contextualization with large language models. *Proceedings of 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-2024)*, 2024, pp. 12406–12410. doi:10.1109/ICASSP48485.2024.10446898
59. **Min Z., Wang J.** Exploring the integration of large language models into automatic speech recognition

systems: An empirical study. *International Conference on Neural Information Processing (ICONIP-2023)*, 2023, pp. 69–84. doi:10.1007/978-981-99-8181-6_6

60. **Chen C., Hu Y., Yang C. H. H., Siniscalchi S. M., Chen P. Y., Chng E. S.** HyPoradise: An open baseline for generative speech recognition with large language models. *Proceedings of the 37th International Conference on Neural Information Processing Systems*, 2023, pp. 31665–31688. doi:10.48550/arXiv.2309.15701
61. **Yang C. H. H., Gu Y., Liu Y. C., Ghosh S., Bulyko I., Stolcke A.** Generative speech recognition error correction with large language models and task-activating prompting. *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU-2023)*, 2023, pp. 1–8. doi:10.1109/ASRU57964.2023.10389673
62. **Maekawa K.** Corpus of spontaneous Japanese: Its design and evaluation. *Proceedings of ISCA/IEEE Workshop on Spontaneous Speech Processing and Recognition*, 2003, paper MMO2.
63. **Nagano T., Kurata G., Thomas S., Kuo H. K. J., Bolanos D., Jung H., Saon G.** LLM based text generation for improved low-resource speech recognition models. *Proceedings of 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-2025)*, 2025, pp. 1–5. doi:10.1109/ICASSP49660.2025.10888566
64. **Shang H., Wang Z., Li J., Liu Y., Zhang Y., Li X.** An end-to-end speech summarization using large language model. *Proceedings of the 25th Annual Conference of the International Speech Communication Association, Interspeech 2024*, 2024, pp. 1950–1954. doi:10.21437/Interspeech.2024-1428
65. **Xu J., Li Y., Wang Z., Zhang Y., Li X., Chen X.** MooER: LLM-based speech recognition and translation models from Moore Threads. *arXiv preprint*, 2024. arXiv:2408.05101. doi:10.48550/arXiv.2408.05101
66. **Nachmani E., Levkovitch A., Hirsch R., Salazar J., Asawaroengchai C., Mariooryad S., Ramanovich M. T.** Spoken question answering and speech continuation using spectrogram-powered LLM. *12th International Conference on Learning Representations (ICLR-2024)*, 2024. doi:10.48550/arXiv.2305.15255
67. **Wang Q., Huang Y., Zhao G., Clark E., Xia W., Liao H.** DiarizationLM: Speaker diarization post-processing with large language models. *Proceedings of the 25th Annual Conference of the International Speech Communication Association, Interspeech 2024*, 2024, pp. 3754–3758. doi:10.21437/Interspeech.2024-2214
68. **Kutsakov A., Maximenko A., Gospodinov G., Bogomolov P., Minkin F.** GigaAM: Efficient self-supervised learner for speech recognition. *arXiv preprint*, 2025. arXiv:2506.01192.
69. **Маслюхин С. М.** Диалоговая система на основе устных разговоров с доступом к неструктурированной базе знаний. *Научно-технический вестник информационных технологий, механики и оптики*, 2023, т. 23, № 1, с. 88–95. doi:10.17586/2226-1494-2023-23-1-88-95

UDC 004.934.2
doi:10.31799/1684-8853-2026-1-19-35
EDN: DSRKFE

Analytical review of the application of large language models for automatic speech recognition

I. S. Kipyatkova[a], PhD, Associate Professor, Senior Researcher, orcid.org/0000-0002-1264-4458, kipyatkova@iias.spb.su
M. D. Dolgushin[a], Junior Researcher, orcid.org/0000-0002-4344-2330
I. A. Kagirov[a], Research Fellow, orcid.org/0000-0003-1196-1117
[a]St. Petersburg Federal Research Center of the Russian Academy of Science, 39, 14th Line, 199178, Saint-Petersburg, Russian Federation

Introduction: One of the trends in natural language processing is the increasing use of large language models. In speech recognition systems, large language models are replacing traditional language models due to their ability to account for broader context. **Purpose:** To systematize and generalize current methods of joint use of automatic speech recognition systems and large language models. **Results:** We identify the main trends in the implementation of large language models to speech recognition. The analysis demonstrates that the application of large language models for hypothesis reranking and error correction consistently improves recognition results, although this improvement is not always fundamental and carries the risk of generating unreliable information due to possible model hallucinations. We conclude that contextualization and in-context learning of large language models can both improve, and degrade recognition results. **Practical relevance:** The generalizations proposed can find practical application in the development of automatic speech recognition systems for various natural and low-resource languages, as well as for code-switched speech. **Discussion:** Recurrent and diffusion large language model architectures have not yet gained widespread use in speech recognition tasks but hold significant potential. A trend towards using decoder-only architectures has been noted, which, in turn, gives rise to the problems of hallucinations and of an orientation towards written norms in text generation.

Keywords — large language models, hypothesis reranking, error correction, in-context learning, automatic speech recognition.

For citation: Kipyatkova I. S., Dolgushin M. D., Kagirov I. A. Analytical review of the application of large language models for automatic speech recognition. *Informatsionno-upravliaiushchie sistemy* [Information and Control Systems], 2026, no. 1, pp. 19–35 (In Russian). doi:10.31799/1684-8853-2026-1-19-35, EDN: DSRKFE

Financial support
This survey was financially supported by budgetary theme No. FFZF-2025-0003.

References

1. Zhao W. X., Zhou K., Li J., Tang T., Wang X., Hou Y., Min Y., Zhang B., Zhang J., Dong Z., Du Y., Yang C., Chen Y., Chen Z., Jiang J., Ren R., Li Y., Tang X., Liu Z., Liu P., Nie J. Y., Wen J. R. A survey of Large Language Models. *arXiv preprint*, 2023. arXiv:2303.18223. doi:10.48550/arXiv.2303.18223
2. Minaee Sh., Mikolov T., Nikzad N., Chenaghlu M. A., Socher R., Amatriain X., Gao J. Large language models: A survey. *arXiv preprint*, 2024. arXiv:2402.06196. doi:10.48550/arXiv.2402.06196
3. Wang F., Zhang Z., Zhang X., Wu Z., Mo T., Lu Q., Wang W., Li R., Xu J., Tang X., He Q., Ma Y., Huang M., Wang S. A comprehensive survey of small language models in the era of large language models: Techniques, enhancements, applications, collaboration with LLMs, and trustworthiness. *arXiv preprint*, 2024. arXiv:2411.03350. doi:10.48550/arXiv.2411.03350
4. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser Ł., Polosukhin I. Attention is all you need. *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS-2017)*, 2017, pp. 6000–6010. doi:10.48550/arXiv.1706.03762
5. Kapusta K. L., Kipyatkova I. S., Kagirov I. A. Analytical survey of transformer-based end-to-end speech recognition models and strategies. *Informatsionno-upravliaiushchie sistemy* [Information and Control Systems], 2024, no. 5, pp. 2–15 (In Russian). doi:10.31799/1684-8853-2024-5-2-15, EDN: MWTGXE
6. Hwang S., Lahoti A., Puduppully R., Dao T., Gu A. Hydra: Bidirectional state space models through generalized matrix mixers. *Proceedings of the 38th Annual Conference on Neural Information Processing Systems (NIPS-2024)*, pp. 110876–110908. doi:10.48550/arXiv.2407.09941
7. Xu W., Hu W., Wu F., Sengamedu S. DeTiME: Diffusion-enhanced topic modeling using encoder-decoder based LLM. *Findings of the Association for Computational Linguistics (EMNLP-2023)*, 2023, pp. 9040–9057. doi:10.18653/v1/2023.findings-emnlp.606
8. Lewis M., Liu Y., Goyal N., Ghazvininejad M., Mohamed A., Levy O., Stoyanov V., Zettlemoyer L. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 7871–7880. doi:10.18653/v1/2020.acl-main.703
9. Liu Y., Gu J., Goyal N., Li X., Edunov S., Ghazvininejad M., Lewis M., Zettlemoyer L. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 2020, vol. 8, pp. 726–742. doi:10.1162/tacl_a_00343
10. Raffel C., Shazeer N., Roberts A., Lee K., Narang Sh., Matena M., Zhou Y., Li W., Liu P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 2020, vol. 21, iss. 1, pp. 5485–5551. doi:10.48550/arXiv.1910.10683
11. Xue L., Constant N., Roberts A., Kale M., Al-Rfou R., Siddhant A., Barua A., Raffel C. mT5: A massively multilingual pre-trained text-to-text transformer. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-2021): Human Language Technologies*, 2021, pp. 483–498. doi:10.18653/v1/2021.naacl-main.41
12. Devlin J., Chang M. W., Lee K., Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-2021*): Human Language Technologies, 2021, pp. 4171–4186. doi:10.18653/v1/N19-1423
13. Kuratov Yu., Arkhipov M. Adaptation of deep bidirectional multilingual transformers for Russian language. *arXiv preprint*, 2019. arXiv:1905.07213. doi:10.48550/arXiv.1905.07213
14. Liu Y., Ott M., Goyal N., Du J., Joshi M., Chen D., Levy O., Lewis M., Zettlemoyer L., Stoyanov V. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint*, 2019. arXiv:1907.11692. doi:10.48550/arXiv.1907.11692
15. Conneau A., Khandelwal K., Goyal N., Chaudhary V., Wenzek G., Guzmán F., Grave E., Ott M., Zettlemoyer L., Stoyanov V. Unsupervised cross-lingual representation learning at scale. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL-2020)*, 2020, pp. 8440–8451. doi:10.18653/v1/2020.acl-main.747
16. Radford A., Narasimhan K., Salimans T., Sutskever I. Improving language understanding by generative pre-training. *OpenAI Blog*, 2018. Available at: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf (accessed 14 May 2025).
17. Chowdhery A., Narang S., Devlin J., Bosma M., et al. *The Journal of Machine Learning Research*, 2023, vol. 24, iss. 1, pp. 11324–11436. doi:10.48550/arXiv.2204.02311
18. Touvron H., Lavril T., Izacard G., Martinet X., Lachaux M. A., Lacroix T., Rozière B., Goyal N., Hambro E., Azhar F., Rodriguez A., Joulin A., Grave E., Lample G. Llama: Open and efficient foundation language models. *arXiv preprint*, 2023. arXiv:2302.13971. doi:10.48550/arXiv.2302.13971
19. Almazrouei E., Alobeidli H., Alshamsi A., Cappelli A., Cojocaru R., Debbah M., Goffinet E., Heslow D., Launay J., Malartic Q., Mazotta D., Noune B., Pannier B., Penedo G. The Falcon series of open language models. *arXiv preprint*, 2023. arXiv:2311.16867. doi:10.48550/arXiv.2311.16867
20. Jiang D., Wu B., Chen C., Li R., Chen G., Sun Y., Kong X., Li L. From clip to dino: Visual encoders shout in multi-modal large language models. *arXiv preprint*, 2023. arXiv:2310.08825. doi:10.48550/arXiv.2310.08825
21. Jiang A. Q., Sablayrolles A., Roux A., Mensch A., Savary B., Bamford C., Chaplot D. S., de las Casas D., Hanna E. B., Bressand F., Lengyel G., Bour G., Lample G., Lavaud L. R., Saulnier L., Lachaux M.-A., Stock P., Subramanian S., Yang S., Antoniak S., Le Scao T., Gervet T., Lavril T., Wang T., Lacroix T., El Sayed W. Mixtral of experts. *arXiv preprint*, 2024. arXiv:2401.04088. doi:10.48550/arXiv.2401.04088
22. Fedus W., Zoph B., Shazeer N. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *The Journal of Machine Learning Research*, 2022, vol. 23, iss. 1, pp. 1–39. doi:10.48550/arXiv.2101.03961
23. Gu A., Dao T. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint*, 2023. arXiv:2312.00752. doi:10.48550/arXiv.2312.00752
24. Peng B., Alcaide E., Anthony Q., Albalak A., Arcadinho S., Biderman S., Cao H., Cheng X., Chung M., Grella M., Kiran G. K., He X., Hou H., Lin J., Kazienko P., Kocon J., Kong J., Koptyra B., Lau H., Mantri K. S. I., Mom F., Saito A., Song G., Tang X., Wang B., Wind J. S., Wozniak S., Zhang R., Zhang Z., Zhao Q., Zhou P., Zhou Q., Zhu J., Zhu R.-J. RWKV: Reinventing RNNs for the transformer era. *Findings of the Association for Computational Linguistics (EMNLP-2023)*, 2023, pp. 14048–14077. doi:10.18653/v1/2023.findings-emnlp.936
25. Beck M., Pöppel K., Spanring M., Auer A., Prudnikova O., Kopp M., Klambauer G., Brandstetter J., Hochreiter S. xLSTM: Extended long short-term memory. *Advances in Neural Information Processing Systems*, 2025, vol. 37, pp. 107547–107603. doi:10.48550/arXiv.2405.04517
26. De S., McLeish T., Botev A., Gu A., Dao T., Goyal N. Griffin: Mixing gated linear recurrences with local attention for efficient language models. *arXiv preprint*, 2024. arXiv:2402.19427. doi:10.48550/arXiv.2402.19427
27. Li X., Thickstun J., Gulrajani I., Liang P. S., Hashimoto T. B. Diffusion-LM improves controllable text generation. *Advances in Neural Information Processing Systems*, 2022, vol. 35, pp. 4328–4343. doi:10.48550/arXiv.2205.14217
28. Nie S., Zhu F., You Z., Zhang X., Ou J., Hu J., Li C. Large language diffusion models. *Proceedings of Workshop on Deep Generative Model in Machine Learning: Theory, Principle and Efficacy (ICLR-2025)*, 2025. doi:10.48550/arXiv.2501.11720
29. Wang M., Liu Zh., Jin Z., Sun G., Zhang Ch., Woodland P. C. Audio-conditioned diffusion LLMs for ASR and deliberation processing. *arXiv preprint*, 2025. arXiv:2509.16622. doi:10.48550/arXiv.2509.16622
30. Shin J., Lee Y., Jung K. Effective sentence scoring method using BERT for speech recognition. *Proceedings of Machine Learning Research*, 2019, pp. 1081–1093. doi:10.48550/arXiv.1910.09932
31. Chiu S. H., Chen B. Innovative BERT-based reranking language models for speech recognition. *IEEE Spoken Language Technology Workshop (SLT-2021)*, 2021, pp. 266–271. doi:10.1109/SLT48900.2021.9383578
32. Shivakumar P. G., Kolehmainen J., Gourav A., Gu Y., Gandhe A., Rastrow A., Bulyko I. Speech recognition rescoring with large speech-text foundation models. *Proceedings of 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-2025)*, 2025, pp. 1–5. doi:10.1109/ICASSP48485.2025.10494321

33. Li S., Ko Y., Ito A. LLM as decoder: Investigating lattice-based speech recognition hypotheses rescoring using LLM. *Proceedings of 2024 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC-2024)*, 2024, pp. 1–5. doi:10.1109/APSIPAASC58517.2024.10373582
34. Ma R., Qian M., Manakul P., Gales M., Knill K. Can generative large language models perform ASR error correction? *arXiv preprint*, 2023. arXiv:2307.04172. doi:10.48550/arXiv.2307.04172
35. Zhao Y., Yang X., Wang J., Gao Y., Yan C., Zhou Y. BART based semantic correction for Mandarin automatic speech recognition system. *Proceedings of the 22nd Annual Conference of the International Speech Communication Association, Interspeech 2021*, 2021, pp. 2017–2021. doi:10.21437/Interspeech.2021-1023
36. Ma R., Gales M. J., Knill K. M., Qian M. N-best T5: Robust ASR error correction using multiple input hypotheses and constrained decoding space. *Proceedings of the 24th Annual Conference of the International Speech Communication Association, Interspeech 2023*, 2023, pp. 3267–3271. doi:10.21437/Interspeech.2023-2189
37. Ma R., Qian M., Gales M., Knill K. ASR error correction using large language models. *IEEE Transactions on Audio, Speech and Language Processing*, 2025, vol. 33, pp. 1389–1401. doi:10.1109/TASLPRO.2025.3551083
38. Li S., Chen C., Kwok C. Y., Chu C., Cheng E. S., Kawai H. Investigating ASR error correction with large language model and multilingual 1-best hypotheses. *Proceedings of the 25th Annual Conference of the International Speech Communication Association, Interspeech 2024*, 2024, pp. 1315–1319. doi:10.21437/Interspeech.2024-368
39. Ko Y., Li S., Yang C. H. H., Kawahara T. Benchmarking Japanese speech recognition on ASR-LLM setups with multi-pass augmented generative error correction. *arXiv preprint*, 2024. arXiv:2408.16180. doi:10.48550/arXiv.2408.16180
40. Wu H., Wang W., Wan Y., Jiao W., Lyu M. ChatGPT or Grammarly? Evaluating ChatGPT on grammatical error correction benchmark. *arXiv preprint*, 2023. arXiv:2303.13648. doi:10.48550/arXiv.2303.13648
41. Mirbeygi M., Beigy H. Prompt guided diffusion for controllable text generation. *Proceedings of the Tenth Workshop on Noisy and User-generated Text*, 2025, pp. 78–84. doi:10.18653/v1/2025.wnut-1.9
42. Li Y., Wang X., Cao S., Zhang Y., Ma L., Xie L. A transcription prompt-based efficient audio large language model for robust speech recognition. *Proceedings of the 25th Annual Conference of the International Speech Communication Association, Interspeech 2024*, 2024, pp. 1905–1909. doi:10.21437/Interspeech.2024-968
43. Yang Z., Chen X., Zhang H., Li Y., Wang Y., Yu D. When large language models meet speech: A survey on integration approaches. *arXiv preprint*, 2025. arXiv:2502.19548. doi:10.48550/arXiv.2502.19548
44. Huang R., Li M., Yang D., Shi J., Chang X., Ye Z., Watanabe S. AudioGPT: Understanding and generating speech, music, sound, and talking head. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024, vol. 38, no. 21, pp. 23802–23804. doi:10.1609/aaai.v38i21.30570
45. Hsu W. N., Bolte B., Tsai Y. H. H., Lakhotia K., Salakhutdinov R., Mohamed A. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2021, vol. 29, pp. 3451–3460. doi:10.1109/TASLP.2021.31222
46. Huang W. C., Chen Z., Chuang P. Y., Harwath D., Glass J. Speech recognition by simply fine-tuning BERT. *arXiv preprint*, 2021. arXiv:2102.00291. doi:10.48550/arXiv.2102.00291
47. Wang M., Han W., Shafran I., Wu Z., Chiu C.-C., Cao Y., Wang Y., Chen N., Zhang Y., Soltau H., Rubenstein P., Zilka L., Yu D., Meng Z., Pundak G., Siddhartha N., Schalkwyk J., Wu Y. SLM: Bridge the thin gap between speech and text foundation models. *Proceedings of 2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU-2023)*, 2023, pp. 1–8. doi:10.1109/ASRU57964.2023.10389703
48. Zhang Y., Qin J., Park D. S., Han W., Chiu C. C., Pang R., Le Q. V., Wu Y. Google USM: Scaling automatic speech recognition beyond 100 languages. *arXiv preprint*, 2023. arXiv:2303.01037. doi:10.48550/arXiv.2303.01037
49. Bai Y., Chen J., Chen J., Chen W., Chen Z., Ding C., Dong L., Dong Q., Du Y., Gao K., Gao L., Guo Y., Han M., Han T., Hu W., Hu X., Hu Y., Hua D., Huang L., Huang M., Huang Y., Jin J., Kong F., Lan Z., Li T., Li X., Li Z., Lin Z., Liu R., Liu S., Lu L., Lu Y., Ma J., Ma S., Pei Y., Shen C., Tan T., Tian X., Tu M., Wang B., Wang H., Wang Y., Wang Y., Xia H., Xia R., Xie S., Xu H., Yang M., Zhang B., Zhang J., Zhang W., Zhang Y., Zhang Y., Zheng Y., Zou M. SEED-ASR: Understanding diverse speech and contexts with LLM-based speech recognition. *arXiv preprint*, 2024. arXiv:2407.04675. doi:10.48550/arXiv.2407.04675
50. Kipyatkova I., Kagirov I., Dolgushin M. Use of pretrained multilingual models for Karelian speech recognition. *Informatics and Automation*, 2025, no. 24(2), pp. 604–630 (In Russian). doi:10.15622/ia.24.2.9
51. Nguyen T., Hoang L. V., Tran H. D. Qwen vs. Gemma integration with Whisper: A comparative study in multilingual SpeechLLM systems Qwen vs. Gemma integration with Whisper: A comparative study in multilingual SpeechLLM systems. *Proceedings of Workshop on Multilingual Conversational Speech Language Model (MLC-SLM)*, 2025. doi:10.21437/MLCSLM.2025-10
52. Zhang D., Li S., Zhang X., Zhan J., Wang P., Zhou Y., Qiu X. SpeechGPT: Empowering large language models with intrinsic cross-modal conversational abilities. *Findings of the Association for Computational Linguistics (EMNLP-2023)*, 2023, pp. 15757–15773. doi:10.18653/v1/2023.findings-emnlp.1055
53. Rubenstein P. K., Asawaroengchai C., Nguyen D. D., Bapna A., Borsos Z., Chaumont Quitry de F., Chen P., El Badawy D., Han W., Kharitonov E., Muckenhirn H., Padfield D., Qin J., Rozenberg D., Sainath T., Schalkwyk J., Sharifi M., Ramanovich T. M., Tagliasacchi M., Tudor A., Velimirović M., Vincent D., Yu J., Wang Y., Zayats V., Zeghidour N., Zhang Y., Zhang Zh., Zilka L., Frank Ch. AudioPaLM: A large language model that can speak and listen. *arXiv preprint*, 2023. arXiv:2306.12925. doi:10.48550/arXiv.2306.12925
54. Anil R., Dai A. M., Firat O., Johnson M., et al. PaLM 2 technical report. *arXiv preprint*, 2023. arXiv:2305.10403. doi:10.48550/arXiv.2305.10403
55. Shon S., Yang C. H. H., Lee H., Kim J., Kim S., Kim T., Lee H. Y. DiscreteSLU: A large language model with self-supervised discrete speech units for spoken language understanding. *Proceedings of the 25th Annual Conference of the International Speech Communication Association, Interspeech 2024*, 2024, pp. 4154–4158. doi:10.21437/Interspeech.2024-1306
56. Du Z., Wang J., Chen Q., Chu Y., Gao Z., Li Z., Zhang S. LauraGPT: Listen, attend, understand, and regenerate audio with GPT. *arXiv preprint*, 2023. arXiv:2310.04673. doi:10.48550/arXiv.2310.04673
57. Zhang F., Geng W., Huang H., Shan Y., Yi C., Qu H. Boosting code-switching ASR with mixture of experts enhanced speech-conditioned LLM. *Proceedings of 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-2025)*, 2025, pp. 1–5. doi:10.1109/ICASSP49660.2025.10890030
58. Lakomkin E., Wu C., Fathullah Y., Kalinli O., Seltzer M. L., Fuegen C. End-to-end speech recognition contextualization with large language models. *Proceedings of 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-2024)*, 2024, pp. 12406–12410. doi:10.1109/ICASSP48485.2024.10446898
59. Min Z., Wang J. Exploring the integration of large language models into automatic speech recognition systems: An empirical study. *International Conference on Neural Information Processing (ICONIP-2023)*, 2023, pp. 69–84. doi:10.1007/978-981-99-8181-6_6
60. Chen C., Hu Y., Yang C. H. H., Siniscalchi S. M., Chen P. Y., Chng E. S. HyPoradise: An open baseline for generative speech recognition with large language models. *Proceedings of the 37th International Conference on Neural Information Processing Systems*, 2023, pp. 31665–31688. doi:10.48550/arXiv.2309.15701
61. Yang C. H. H., Gu Y., Liu Y. C., Ghosh S., Bulyko I., Stolcke A. Generative speech recognition error correction with large language models and task-activating prompting. *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU-2023)*, 2023, pp. 1–8. doi:10.1109/ASRU57964.2023.10389673
62. Maekawa K. Corpus of spontaneous Japanese: Its design and evaluation. *Proceedings of ISCA/IEEE Workshop on Spontaneous Speech Processing and Recognition*, 2003, paper MMO2.

63. Nagano T., Kurata G., Thomas S., Kuo H. K. J., Bolanos D., Jung H., Saon G. LLM based text generation for improved low-resource speech recognition models. *Proceedings of 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-2025)*, 2025, pp. 1–5. doi:10.1109/ICASSP49660.2025.10888566
64. Shang H., Wang Z., Li J., Liu Y., Zhang Y., Li X. An end-to-end speech summarization using large language model. *Proceedings of the 25th Annual Conference of the International Speech Communication Association, Interspeech 2024*, 2024, pp. 1950–1954. doi:10.21437/Interspeech.2024-1428
65. Xu J., Li Y., Wang Z., Zhang Y., Li X., Chen X. MooER: LLM-based speech recognition and translation models from Moore Threads. *arXiv preprint*, 2024. arXiv:2408.05101. doi:10.48550/arXiv.2408.05101
66. Nachmani E., Levkovitch A., Hirsch R., Salazar J., Asawaroengchai C., Mariooryad S., Ramanovich M. T. Spoken question answering and speech continuation using spectrogram-powered LLM. *12th International Conference on Learning Representations (ICLR-2024)*, 2024. doi:10.48550/arXiv.2305.15255
67. Wang Q., Huang Y., Zhao G., Clark E., Xia W., Liao H. DiarizationLM: Speaker diarization post-processing with large language models. *Proceedings of the 25th Annual Conference of the International Speech Communication Association, Interspeech 2024*, 2024, pp. 3754–3758. doi:10.21437/Interspeech.2024-2214
68. Kutsakov A., Maximenko A., Gospodinov G., Bogomolov P., Minkin F. GigaAM: Efficient self-supervised learner for speech recognition. *arXiv preprint*, 2025. arXiv:2506.01192.
69. Masliukhin S. M. Dialogue system based on spoken conversations with access to an unstructured knowledge base. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2023, vol. 23, no. 1, pp. 88–95 (In Russian). doi:10.17586/2226-1494-2023-23-1-88-95